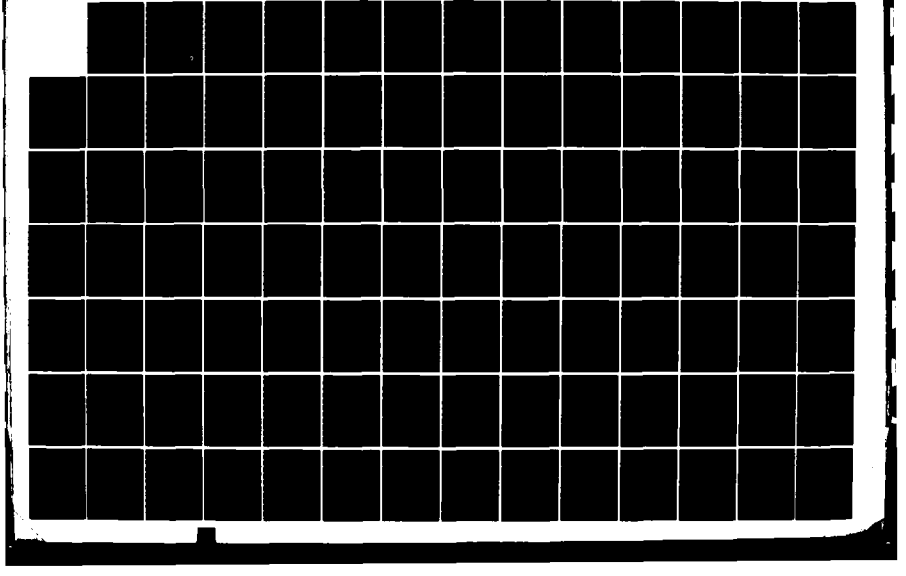
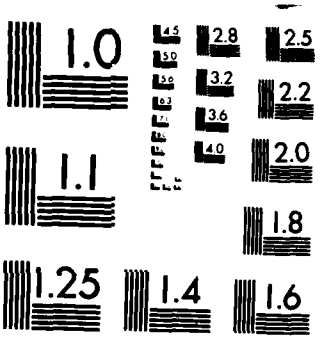
AD-A151 877 PREDICTING COLLEGE DROPOUTS BY COMBINING AUTOMATIC INTERACTION DETECTOR A.. (U) AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH S R SCHMIDT DEC 84 1/2

UNCLASSIFIED AFIT/CI/NR-85-25D F/G 12/1 NL





UNCLASS
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFIT/CI/NR 85-25D	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Predicting College Dropouts By Combining Automatic Interaction Detector And Discriminant Analysis	5. TYPE OF REPORT & PERIOD COVERED: THESIS/DISSERTATION
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Stephen R. Schmidt	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: AFIT STUDENT AT: University of Northern Colorado	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: AFIT/NR, WPAFB OH 45433	12. REPORT DATE: Dec 1984
	13. NUMBER OF PAGES: 98
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): AD-A151 877	15. SECURITY CLASS. (of this report): UNCLASS
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

APPROVED FOR PUBLIC RELEASE: IAW AFR 190-1

LYNN E. WOLAVER 28 Feb 85
Dean for Research and
Professional Development
AFIT, Wright-Patterson AFB OH

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

ATTACHED

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE UNCLASS
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT

SCHMIDT, Stephen R. "Predicting College Dropouts By Combining Automatic Interaction Detector and Discriminant Analysis."

A two group discriminant analysis was performed on two large samples from the Air Force Academy to predict college success and failure. The efficiency of the model was estimated by the hit rate (i.e. the proportion of correctly classified subjects) and by a cross-validation process in which difference in hit rates (shrinkage) was calculated. A new procedure, MAIDDA, was developed which combines a modified automatic interaction detector (MAID) with two group discriminant analysis. The MAID procedure does not require the conversion of continuous variables to categorical variables. In addition, MAID is easily performed with existing statistical software packages. The unique contribution of MAID to the prediction process was estimated by differences in hit rates and shrinkage for the two group discriminant analysis and MAIDDA when applied to the same sample data. The results from the samples described above indicate a substantial improvement in prediction when MAIDDA is used. It is postulated that MAIDDA will provide prediction improvement for most samples where $N \geq 1000$, numerous

25

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

PREDICTING COLLEGE DROPOUTS BY COMBINING AUTOMATIC INTERACTION DETECTOR AND DISCRIMINANT ANALYSIS

A Dissertation Submitted in Partial Fulfillment
of the Requirement for the Degree of
Doctor of Philosophy

Stephen R. Schmidt

College of Arts and Sciences
Department of Mathematics and Applied Statistics

December, 1984

THIS DISSERTATION WAS SPONSORED

BY

Samuel R. Houston, Ph.D
(Research Advisor)

Stephen R. Schmidt

DISSERTATION COMMITTEE

Advisory Professor _____
Donald T. Searls, Ph.D.

Advisory Professor _____
Dale Shaw, Ph.D.

Faculty Representative _____
Garth Eldredge, Ph.D.

DEAN OF THE GRADUATE SCHOOL

_____

Examination Date of Dissertation _____

ACKNOWLEDGMENTS

I am very grateful for the support, love and understanding that I received from my wife, Judy, and my son, Jeff, over the last 26 months. They are both exceptional people whom I love dearly and without them I would have had difficulty completing this program. I especially thank Judy for her patience despite the numerous times she retyped this paper. In addition, I am indebted to my parents and friends for their encouragement and their prayers.

Most importantly, I thank God, my Creator, for answering our prayers and for causing all things to work together for good. For many reasons, this has been an unforgetable 26 months and without God's love and guidance, I would not have made it.

TABLE OF CONTENTS

predictors are available and interaction of predictor variables exists. Further tests of MAIDDA are needed to ascertain its full potential.

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF FIGURES

CHAPTER I

INTRODUCTION

For more than fifty years, administrators of higher education have struggled with the problem of student attrition. Researchers have attempted to investigate this problem by developing statistical designs using multiple regression and discriminant models to predict accurately a potential student's success or failure. Most of the earlier studies claim to account for 60% to 70% of the variance in persister/dropout models. In recent attempts to enhance model accuracy, some researchers have developed more complex designs such as maximum likelihood and log-linear models; however, these methods have not produced substantial improvements. One possible cause for this lack of success is the assumption that persister/dropout models are additive, i.e., interaction does not exist. The area of nonadditivity has not been fully explored due to the difficulty in developing appropriate interaction terms when there are numerous predictor variables. For this reason, most researchers conveniently assume an additive model to be appropriate.

A relatively new procedure that automatically accounts for the interaction terms in a prediction model is the Automatic Interaction Detector, referred to as AID (Sonquist, Baker and Morgan, 1971). Thus far, AID has been applied as an independent procedure, and when significant interaction is present, results have surpassed those of multiple regression (Karathanos, 1975). Since multiple regression and discriminant analysis do not easily handle complex nonadditive models, it appears that combining either of these procedures with AID has the potential of enhancing prediction.

Another topic to be addressed in this study is how to measure persister/dropout model effectiveness. A reasonable approach to this problem presented by Allen and Yen (1979) is the hit rate, i.e., the proportion of correct predictions. This procedure has been reported only once in persister/dropout studies, indicating that previous researchers appear so concerned about correlation coefficients that they actually failed to perform individual subject predictions. Model comparisons between χ^2 values are confusing due to the dependence of this statistic on the sample size. For example, consider a 2 x 2 contingency table where N equals the total sample size. The maximum χ^2 value is N which occurs for either 100% correct prediction or 100% incorrect prediction. For any two 2 x 2 contingency tables with sample sizes N_I and N_{II} and matching cell proportions $\chi^2_I / N_I = \chi^2_{II} / N_{II}$. The statistic $\phi = \sqrt{\chi^2/N}$ compensates for

different N_j values; however, problems with using ϕ are pointed out in the Chapter 5 summary. The squared correlation coefficient, R^2, also presents problems when comparing models on different sample sizes and/or different numbers of predictor variables (Marasciulo and Levin, 1983). On the other hand, between studies of different size samples, hit rates can be compared with less ambiguity. A further measure of model efficiency and generalization is determined through a cross-validation where the drop in hit rate is referred to as the amount of shrinkage in the model.

Purpose of the Study

The purpose of the study is to explore a new prediction process that can automatically account for variable interaction and enhance prediction for large samples with numerous predictor variables. The development of the new process includes: (a) a modification of AID that will enhance efficiency and allow for execution through the use of current statistical software packages, (b) a combination of the modified AID procedure with discriminant analysis, and (c) the implementation of the hit rate as a measure of model efficiency.

Signficance of the Study

There are five main areas where this study is attempting to make a significant contribution to the

advancement of persister/dropout models and the field of statistics. First, researchers attempting to develop prediction models for college dropouts currently face numerous alternatives to methodological procedures, some of which are extremely complex; most of these methods have produced less than desirable results. The literature review for the study will aid future researchers in eliminating unnecessary complex designs and point out method similarities which lead to similar results.

Second, a new procedure combining two group discriminant analysis and a modified automatic interaction detector procedure (MAID) is presented as a means of handling nonadditivity and enhancing prediction. Through the use of two large samples, the effectiveness of this new procedure and the unique contribution of MAID will be presented.

Third, the MAID procedure to be incorporated is modified from that of Sonquist et al. (1971) in order that continuous variables need not be converted to categorical data. This in itself provides a possible advancement in the use of automatic interaction detectors.

Fourth, the failure of many previous persister/dropout studies to report the hit rate as a measure of model efficiency is disturbing. It is hoped that this study will encourage future researchers to conduct individual subject predictions and to report hit rates that will allow for unambiguous model comparisons.

The fifth area of emphasis is to demonstrate the usefulness of a prediction model in an actual institutional setting. Since the data utilized in this study will be obtained from cadets at the U.S. Air Force Academy, any increase in predictability of college success or failure will aid the Academy in determining future admissions criteria which can ultimately reduce attrition. The Academy was founded in 1955 and the attrition rate has fluctuated between 24% and 46% with the current dropout rate reported to be approximately 38% (Jensen, 1983). Each cadet who drops out costs the taxpayers of this country an average of $25,000. Therefore, a substantial reduction in attrition, i.e., a drop from 38% to 28% over the four year program could save over three million dollars a year, at the Air Force Academy alone. Although this study is aimed specifically at the Air Force Academy, the same procedures should be applicable at the other service academies or at civilian institutions of higher education.

Terminology

The data used in this study will be from two large samples of cadets at the U.S. Air Force Academy. In order to avoid confusion concerning the terminology unique to the Academy, the following word list and definitions are presented (USAFA Catalog, 1981):

Applicant - An individual who applies to a member of Congress or another nominating authority requesting a nomination.

Nomination - The result of naming an applicant as an academy candidate by a nominating authority.

Nominee - An applicant who has obtained a nomination in a category authorized by law.

Candidate - A legally nominated individual whose name has been recorded by the Director of Cadet Admissions.

Appointment - An offer of admission to a fully qualified candidate.

Appointee - A qualified candidate who has been selected for admission.

Cadet - Student enrolled at the Academy.

Turnback - A cadet who has been turned back to a subsequent class.

Dropout - A cadet who is permanently disenrolled prior to completing the four year program, i.e., a voluntary or nonvoluntary permanent withdrawal.

Persister - A cadet who completes the program in the normal 48-month period.

Stopout - A cadet who has left the Academy for up to one year and then returned to resume his or her program.

Fourth class cadet - freshman

Third class cadet - sophomore

Second class cadet - junior

First class cadet - senior

Cadet Wing - Student body of cadets which is limited by law to approximately 4400.

Class of 19XX - a cadet class which enters together and is scheduled to graduate in 19XX.

Military Performance Average (MPA) - average military grade similar to GPA in that possible scores are 0 = F, 1 = D, 2 = C, 3 = B, and 4 = A. Grades are based on leadership potential as evaluated by their commanders, faculty and peers.

Physical Aptitude Examination (PAE) - combined score based on performance of several physical events, such as pull-ups, standing broad jump, modified basketball throw, agility run, and a 300-yard shuttle run.

Military Order of Merit (MOM) - former system of rank ordering of cadets based on leadership potential.

CHAPTER II

SELECTED REVIEW OF RELATED LITERATURE

While the purpose of this study pertains to predicting persisters and dropouts, the literature review will emphasize prediction versus explanatory models. It is for this reason that most of the references are before 1975 when prediction studies were prevalent. Since that time, the majority of the emphasis has been on explanatory models, which are not discussed at length in this report. A general review of the literature is provided to bring the reader up to date with the persister/dropout problem and to review the types of variables that are typically included in a prediction model. A separate section on Air Force Academy literature is also presented because the data for this study are obtained from Academy cadets. The most important part of this chapter pertains to the methodological review that provides a basis for the development of a new procedure that will, hopefully, enhance the prediction of persisters and dropouts and, thereby, advance the knowledge in this area of statistics.

General Review of Literature

Early attempts to analyze the relationships between college success and a set of predictor variables utilized college grade point average (GPA) as the criterion variable with one predictor at a time. The primary noncognitive predictor variables studied were participation in extra-curricular activities (Twining, 1957), study habits (Chahbazi, 1957), attitude (Myers and Schultz, 1950), motivation (DiVesta et al; 1949), biographical inventory (Anastasi, Schneider, and Meade, 1960), age and gender (Summerskill, 1962), parents' occupation and education (Bonner, 1956).

According to several studies listed in Table 1, the most successful analyses were done with cognitive measures such as high school academic record and college aptitude tests. In Table 1 is found the simple correlation coefficients between the cognitive predictors and college academic achievement for several studies. Of all the previously mentioned variables, high school academic record was found to be the best overall predictor for college freshman GPA, i.e., it consistantly had the highest simple correlation coefficient with college achievement (Berdie, 1962; Gallant, 1965).

Due to the inefficiency of the simple regression models, numerous studies were conducted with multiple correlations including different combinations of two

Table 1

Successful Predictors of College Academic Achievement and Their Simple Correlation Coefficients

Predictor	Correlation Coefficent	Reference
High School Achievement	.55	(Garrett, 1949)
	.55	(Fricke, 1956)
	.50	(Fishman and Pasanella, 1960)
	.63	(Gallant, 1965)
	.40	(Michael and Jones, 1962)
	.40	(Webb and McCall, 1963)
High School Rank	.55	(Garrett, 1949)
	.50	(Fishman and Pasanella, 1960)
	.64	(Berdie, 1962)
	.60	(Bonner, 1956)
College Entrance Exams in Math	.49	(Bonner, 1956)
	.61	(Webb and McCall, 1963)
	.47	(Boyce and Paxson, 1965)
	.49	(Bonner, 1956)
College Entrance Exams in English	.56	(Webb and McCall, 1963)
	.64	(Boyce and Paxson, 1965)
	.47	(Bonner, 1956)
	.51	(Boyer, 1956)
College Entrance Exams in Social Studies	.50	(Boyce and Paxson, 1965)
College Entrance Exams in Natural Sciences	.46	(Boyce and Paxson, 1965)

predictor variables. The range of multiple correlation coefficients was found to be .54 to .81 with a median of about .57 (Garrett, 1949). In 216 studies by Fishman and Pasanella (1960), multiple correlations ranged from .37 to .83 with a median of .62. The predictors were cognitive measures such as the Scholastic Aptitude Test (SAT), College Entrance Examination Board (CEEB), American Council on Education Psychological Examination for College Freshman (ACE), and high school record. The usual two predictor combination is an aptitude test and high school record, where the criterion is freshman year grade average. Fricke (1956) reported multiple correlations ranged from .60 to .65 when using high school achievement and college board tests, but contradictions to Fricke's results, as well as almost all conclusions reported, can be found in the voluminous studies. In fact, Endler and Steinberg (1963) concluded the only thing consistent in the literature on predicting academic achievement is inconsistency. One of the reasons for this lack of consistency is that many studies used only one variable or just a few variables at a time in the predictive model without controlling other relevant variables. Lavin (1967) also reported that many researchers were predicting GPA in specific areas of study (i.e. engineering, business, etc.) while others tried to predict overall GPA. With this in mind, Panos and Astin (1968), in their review of the literature, argue that many differences in research findings existed because the researchers were in fact

dealing with different phenomena (i.e. different sets of predictors and different criterion variables). The lack of complete longitudinal studies also adds to the confusion. Some studies predicted first semester GPA, freshman year GPA, or GPA by category: freshman, sophomore, junior, or senior. Lavin (1967), in his recommendations for future research, supports longitudinal data to cover all four years in the same study.

Numerous studies on college dropouts were conducted from the mid 1960s to the mid 1970s. This time frame coincides with an increased national interest in education and more available government funding for research. Most of these studies tend to support previous findings as to predicting college success as measured by freshmen GPA. In addition, the GPA for college freshmen was found to correlate with whether or not a student persisted to graduation (Astin, 1971) as seen in Figure 1. Thus, it appears that the better predictors for GPA should also be associated with persistance.

Continued studies on predictor variables revealed that for extreme groups (i.e., highly intelligent or well below-average), gender was found to be a significant predictor, and separate analysis for males and females should be accomplished due to possible interaction effects (Lavin, 1967). An individual's vocational goals, such as those measured by the Strong Vocational Interest Blank, were found to be good predictors of college success (Pantages and

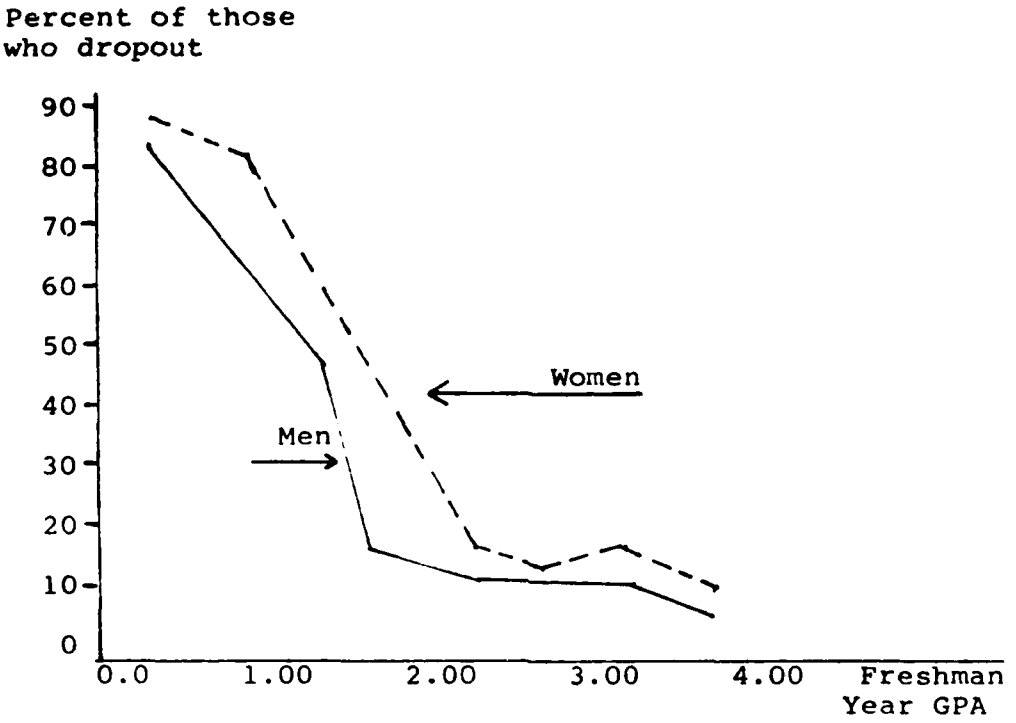


Figure 1. Relationship Between Freshman GPA and Dropping Out of College (N = 36,581 students)

Creedon, 1978). Other individual characteristics, such as the student's race, parents' education and economic status were noted as contributing only a small amount to the prediction model (Astin, 1971). However, some studies report race as a significant predictor indicating that maybe the problem is not with the race variable but with the way it is coded. For example, Smith (1982) concluded race was not a significant predictor in his two group discriminant analysis; however, he chose to code race as follows: 1, if American Indian; 2, if Black, 3, if Oriental; 4, if Spanish; 5, if Caucasian; and 6, if other. This type of coding assumes an underlying continuum of a variable that is at the

nominal level of measurement. Pascarella and others (1981) found that race, when binary coded is a significant predictor in discriminant analysis. It is, therefore, critical for a researcher to investigate the proper coding for nominal variables representing multiple categories.

In summary, the primary variables that are typically included in predicting persisters and droputs include: high school academic record; high school rank; college aptitude exams, such as SAT and ACT; and gender. Interaction of the predictor variables has not been studied extensively; however, Pascarella and Chapman (1983) have shown that interaction of some non-cognitive variables may be significant. Lavin (1967) highly suspected interaction effects and recommended they be studied. In addition to the variable information described above, it was also noted that most studies tend to estimate the success of their model by the value of the correlation coefficient. However, it is difficult to measure or compare the usefulness of any reported models without the actual prediction process and the determination of a hit rate. With the national college attrition rate at about 40%, (Summerkill, 1962), the prediction of success for all subjects would provide a base rate of .60. New methods must improve on this rate to be efficient. Some studies have reported χ^2 and ϕ values as measures of model efficiency; however, these values, as well as correlation coefficients, are dependent on the sample size and there is not a functional relationship between them and the hit rate.

Therefore, it is important that the hit rate be reported to estimate adequately model efficiency.

Review of U.S. Air Force Academy Studies

In the summer of 1955, the first class of 306 cadets entered the United States Air Force Academy. Since then the Academy Cadet Wing has grown to over 4000. During the four year program, which leads to a Bachelor of Science degree, the Academy has experienced attrition rates ranging from 24% to 46% with the current rate at about 38% (Jensen, 1983). The majority of the students who dropout do so during the first year for voluntary reasons (i.e., dissatisfaction with the program or change of career goals). In many cases, however, the cadet dropout exhibits poor academic or military performance or both. Therefore, it is difficult to differentiate the voluntary dropout from one who is dismissed for academic or military reasons. Many cadets would rather quit voluntarily than be dismissed; others, who attend USAFA only due to parental pressure, feel uneasy about quitting and, therefore, purposely perform poorly in order to get dismissed. For this reason, it appears inappropriate to develop separate prediction models for voluntary and involuntary dropouts.

The previous studies for predicting success at the Academy concentrated on first year GPA as the criterion variable (Miller, 1968; Jernigan, 1969). The findings in

Jernigan's report are not all that clear and since a cross-validation was not performed, there is no measure of external validity. In addition, Jernigan did subgroup analysis for three groups: (a) preparatory school graduates; (b) recognized athletes; and (c) other, which might possibly restrict the predictor variable range within groups, and thereby, reduce efficiency. The most important contribution of Jernigan's study is the demonstration of significant differences in these three groups. Therefore, future researchers should consider including these groups as binary coded predictor variables for future studies.

Because of the increase in the number of dropouts during the early 1970s, the Department of Defense was required to provide a Report to Congress (1976) on student attrition at the four federal service academies. In this report, several predictor variables or characteristics were analyzed and percentages of dropouts were examined for each characteristic. The study was longitudinal in nature and utilized a subgroup expectancy table that could predict success or failure; however, individual predictions were not attempted. The conclusions of this report regarding significant contributions of independent variables are listed as follows:

a. SAT math scores were highly significant for fourth class cadets;

b. High school achievement (i.e., high school rank, grades and honor society membership) are highly significant;

c. During the third class year verbal test scores are significant;

d. Military Order of Merit (MOM) was significant;

e. The Physical Aptitude Exam (PAE) was significant;

f. The athletic activity index, based on participation in high school or community sponsored sports programs had no significance.

g. Non-athletic activities index (i.e., president of high school organizations, major part in a play, edited or worked on a school paper, participated in a state or regional speech or debate) was not significant;

h. Parental income and formal education were not significant;

i. Socioeconomic status was not significant.

The report does not present a cross-validation or the proportion of cadets that can be correctly classified. Therefore, it is more of a descriptive study lacking well defined predictive procedures.

A mathematical equation was developed by Dempsey and Fast (1976) using a dichotomous dependent variable and a maximum likelihood technique to predict dropouts based on admissions data. They studied the Class of 1977 to develop a model for predicting only voluntary dropouts during the first year, and they cross-validated the model on the Class of 1979. This is the only clearly presented model discovered in the Academy literature which predicts on an individual basis whether or not the student will persist or

dropout. They were able to classify correctly 32.1% of the voluntary dropouts and 94.2% of the persisters for the Class of 1977 (see Table 2). The variables found to be significant predictors were the binary coded Strong Variable Index

Table 2

Class of 1977 (N=1183 cadets) Prediction Results

	ACTUAL Dropouts	Persisters	TOTAL
PREDICTED Dropouts	79	55	134
PREDICTED Persisters	167	882	1049
TOTAL	246	937	1183
Percent correctly classified	32.1%	94.2%	81.23%

Blank (SVIB) items relating to interest in math activities, interest in science, and interest in military activites. The cross-validation was performed on the Class of 1979, using only the first semester data, which resulted in correctly identifying 36% of the voluntary dropouts and 91.3% of the persisters (see Table 3). When considering only voluntary dropouts, the attrition for the Class of 1977 is 20.8%. Therefore, the base rate needed to be improved on is 1-.208 = .792. The hit rate for Dempsey and Fast's (1976) procedure was .8123 which is a 3% improvement. In

Table 3

Class of 1979 Prediction Results (N=1460 cadets)

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	64	110	174
PREDICTED			
Persisters	114	1147	1251
TOTAL	178	1257	1460
Percent Correctly Classified	36.0%	91.3%	82.94%

terms of reducing an overall attrition rate of 38.7% for the Class of 1979, the model described above predicted 174 dropouts, thus eliminating 110 successful candidates. So, replacing these 174 candidates with other qualified candidates who are predicted with (1147/1251)(100%) = 91.6% accuracy will reduce the overall number of dropouts from 568 to 519. The revised attrition will be 35.55%. Therefore, the 3% improvement of the hit rate is associated with an approximate 3.2% reduction in attrition. These results remain somewhat ambiguous, though, due to the study's limitation of predicting only first year voluntary dropouts and then projecting a revised attrition over a four year period. It should be noted that a recent unpublished study with data from the Class of 1985 was made by the USAFA Institutional Research Branch comparing a stepwise regression method with

the maximum likelihood procedure. The results, shown in Table 4, do not indicate a substantial difference in the predictability of these procedures (Jensen, 1983); however, the hit rates for both procedures are substantially less than those for the 1977 and 1979 classes. This may indicate a lack of predictive consistency from one time frame to another and also the importance of the SVIB variables which are not included in more recent analysis.

These studies of Academy dropouts appear to be limited because, beginning with the Class of 1980, women have been admitted to the Academy, and no study to include gender has not been reported. In addition, there has not been a complete study of possible interaction effects, and neither Jernigan (1964) nor Dempsey and Fast (1976) examine the predictability of race as an independent variable.

Review of Methodologies

Many of the early studies, such as Summerskill (1962), were descriptive in nature; however, some researchers such as Garrett (1949) used simple correlation coefficients to determine predictor effectiveness. Other studies relied on expectancy tables (Astin, 1971; Report to Congress, 1976) to describe dropouts and to predict success or failure. In the expectancy table models, subgrouping was performed to obtain mutually exclusive cells of homogeneous subjects. Since most studies used only a few selected predictor variables

Table 4

Comparison of the maximum likelihood and multiple Regression Methods using the Class of 1985

MAXIMUM LIKELIHOOD MODEL

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	79	133	212
PREDICTED			
Persisters	182	701	883
TOTAL	261	834	1095
Percent Correctly Classified	30.3%	84.1%	71.2%

MULTIPLE REGRESSION MODEL

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	78	138	216
PREDICTED			
Persisters	183	696	879
TOTAL	261	834	1095
Percent Correctly Classified	29.9%	83.5%	70.7%

for subgrouping without an interaction investigation, it is possible that prediction efficiency was weakened due to the restricted range of the variables within each subgroup (Pedhazer, 1982). Attempts to improve R^2 values led

researchers such as Fishman and Pasanella (1960) and Fricke (1956) to seek multiple correlations. Most of these multiple regression studies analyzed a few predictors at a time with limited success. The development of highly efficient computers made possible the technique of including numerous variables, as in a study by Panos and Astin (1968) which used 20 predictors. Despite these attempts at improving predictability, some researchers remained skeptical over the use of a dichotomus criterion variable and the Bernoulli nature of the error term (Nerlove and Press, 1973). This concern led to a procedure incorporating maximum likelihood and Marshallian Utility Theory used by Dempsey and Fast (1976). Their procedure is more complex than a multiple regression model, but a consistant pattern of improved predictions has not been demonstrated (Dempsey and Fast, 1976; Jensen, 1983). Continued empirical analysis of dichotomous criterion has revealed that when the proportion of 1's for the binary coded variable were between .25 and .75 and sometimes as extreme as .1 to .9, the multiple regression, log-linear and maximum likelihood methods are approximately the same (Knoke 1975; Dempsey and Fast, 1976; Goodman 1976). A more recent study has shown that when the number of predictor variables is large (i.e., greater than 10) and the sample size is much larger still (i.e., greater than 40), then the predicted values for the dependent variable are approximately normally distributed for each level, and, thus, the normality assumption

appears to be satisfied (Marascuilo and Levin, 1983). These results have demonstrated that multiple regression related techniques remain a suitable methodology for future analyses.

The multiple regression procedure develops a prediction equation, $\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 \ldots + b_p x_p$ where $\hat{Y}$ is the predicted value of the dependent variable Y, x_i are the independent or predictor variables, and b_i are the raw score regression weights. Thus, $\hat{Y}$ is a linear combination of the predictor variables where the regression weights are determined by minimizing the sum of squared errors in prediction, i.e.,

$$\text{Min}\left[\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2\right] \quad \text{(Pedhazer, 1982).}$$

A similar procedure is two group discriminant analysis which produces a linear discriminant function $L = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p$. Calculation of the discriminant weights, $\underline{A} = (a_1, a_2, a_3, \ldots, a_p)$ was shown by Sir R. A. Fisher in 1936 to be $\underline{A} = \underline{S}^{-1} * (\underline{\bar{X}}_I - \underline{\bar{X}}_{II})$ where $\underline{S}^{-1}$ is the inverse of the pooled variance-covariance matrix for the two groups and $(\underline{\bar{X}}_I - \underline{\bar{X}}_{II})$ is a p by 1 vector of group mean differences. In this case, the set of raw score discriminant weights are found in such a way that the difference in group means, $\bar{L}_I - \bar{L}_{II}$, is maximized. The larger this value, the more disciminatory power is available and the greater the prediction efficiency. This procedure is the same as maximizing $(\bar{\hat{Y}}_I - \bar{\hat{Y}}_{II}) / S_p$ in the multiple regression model where S_p is the square root of the pooled

variance of the two groups of predicted scores (Marascuilo and Levin, 1983).

Although seemingly different procedures, multiple regression and two group discriminant analysis develop sets of raw score weights for the predictor variables that are proportional, (i.e., $b_i = ka_i$ for i = 1, 2, ... p and some constant k (Michael and Perry, 1956)). However, b_0 is not necessarily equal to ka_0. Therefore, for any subject, i, $(Y - b_0) = k(L - a_0)$ where b_0 and a_0 are the intercept constants of the two linear functions. This relationship of the two methods implies that any decision rule which assigns subject i to group 1 or 2 by way of the corresponding discriminant score, L_i , is also applicable to the predicted score, $\hat{Y}_i$, in multiple regression. In Marascuilo and Levin (1983) the following decision rules for classifying subjects are presented:

Rule 1: Determine which group has the higher group mean and assuming that this is group j, assign the first N_j subjects with the highest discriminant or predicted scores to group j. Assign all other subjects to the other group.

Rule 2: Determine a cut-off score, c, such that

$$c_L = \frac{N_I \bar{L}_I + N_{II} \bar{L}_{II}}{N_I + N_{II}} \quad \text{or} \quad c_{\hat{Y}} = \frac{N_I \bar{\hat{Y}}_I + N_{II} \bar{\hat{Y}}_{II}}{N_I + N_{II}}$$

and assign all subjects with scores greater than c to the group with the larger mean.

Rule 3: When p is large and N is even larger (i.e., $p \geq 10$ and $N \geq 40$) there is reason to believe L and $\hat{Y}$ are distributed approximately normal for each group (see Figure 2).

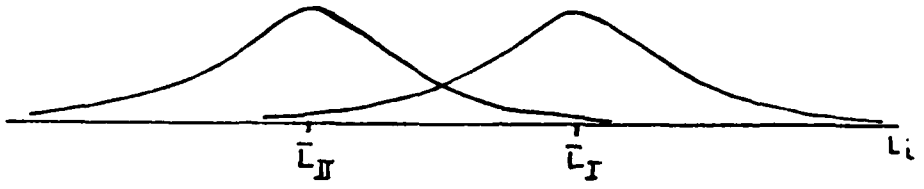


Figure 2. Distribution of Discriminant Scores for Subjects in Groups I and II

If $N_I = N_{II}$, the type I and type II errors can be equalized by setting the cut-off score at

$$C = \frac{\frac{\bar{L}_I}{S_{L_I}} + \frac{\bar{L}_{II}}{S_{L_{II}}}}{\frac{1}{S_{L_I}} + \frac{1}{S_{L_{II}}}}$$ and similarly for $C_{\hat{y}}$.

Rule 4: When $N_I \neq N_{II}$, Rule 3 can be modified as follows:

$$C = \frac{\frac{\bar{L}_I}{S_{L_I}} + \frac{\bar{L}_{II}}{S_{L_{II}}}}{\frac{N_I}{S_{L_I}} + \frac{N_{II}}{S_{L_{II}}}}$$ and similarly for $C_{\hat{y}}$.

Rule 5: This rule is also based on large p and N such that L and $\hat{Y}$ are approximately normally distributed for each group. In addition, define

$$Z_I = \frac{L - \bar{L}_I}{S_{L_I}} \quad \text{and} \quad Z_{II} = \frac{L - \bar{L}_{II}}{S_{L_{II}}}.$$

Therefore, $f(Z_j) = (1/2\pi)^{1/2} EXP(-.5*Z_j^2)$, j = I, II such that $f(Z_I)$ and $f(Z_{II})$ are the corresponding likelihoods. The value associated with the likelihood $f(Z_j)$ is equal to the height of the normal curve at Z_j. Since these values are proportional to probabilities, the conditional probabilities are defined as:

$$P_I = \frac{f(Z_I)}{f(Z_I)+f(Z_{II})} \text{ and } P_{II} = \frac{f(Z_{II})}{f(Z_I)+f(Z_{II})} .$$

In this case, P_I is the probability that subject i with a score L is in group 1 and P_{II} is the probability that the same subject i is in group 2. Subject i is assigned to the group j with the larger P_j. If the variance for groups 1 and 2 are approximately equal, replace S_{L_I} and $S_{L_{II}}$ with S_p, which is the square root of the pooled variance:

$$S = \sqrt{MSW} = \sqrt{\frac{(n_I-1)S_{L_I}^2 + (n_{II}-1)S_{L_{II}}^2}{n_I+n_{II}-2}}$$

Again, this same rule can be applied for multiple regression scores, $\hat{Y}$.

It is also possible to expand Rule 5 to incorporate a decision rule which takes into consideration prior or Bayesian probabilities (Birnbaum and Maxwell, 1960). Referring to Hayes (1963), the appropriate posterior probabilities, PB_j , j = I, II, are

$$PB_I = \frac{PR_I * P_I}{PR_I * P_I + PR_{II} * P_{II}} \text{ and } PB_{II} = \frac{PR_{II} * P_{II}}{PR_I * P_I + PR_{II} * P_{II}}$$

where PR_j is the prior probability of being in group j. As before, assign subject i to the group j which has the larger PB_j.

Tatsuoka (1971) presents a decision rule which is similar to the expanded Rule 5 above. However, instead of using the linear discriminant scores, L_i, Tatsuoka's procedure is a function of the standardized squared distance from group centroids. This procedure assumes that the density function associated with each group of subjects evaluated on p variables has a p variate normal distribution. The likelihood function for a specific subject i given group j is $f(\underline{X}_{ij}, \underline{S}_j)$ where $\underline{X}_{ij} = (x_{i1} - \bar{x}_{j1}, x_{i2} - \bar{x}_{j2}, \ldots, x_{ip} - \bar{x}_{jp})$ and $\underline{S}_j$ is the group j variance-covariance matrix of the p predictors. The equation for the likelihood function is

$$f(\underline{X}_{ij}, \underline{S}_j) = \left(\frac{1}{2\pi}\right)^{p/2} * |\underline{S}_j|^{-1/2} * EXP\left[(-.5)(\underline{X}_{ij}^T * \underline{S}_j^{-1} * \underline{X}_{ij})\right].$$

The conditional probability of subject i being in group j = I, II is represented by

$$P_j = \frac{f(\underline{X}_{ij}, \underline{S}_j)}{\sum_{k=I,II} f(\underline{X}_{ik}, \underline{S}_k)} = \frac{|\underline{S}_j|^{-1/2} * EXP\left[(-.5)(\underline{X}_{ij}^T * \underline{S}_j^{-1} * \underline{X}_{ij})\right]}{\sum_{k=I,II} |\underline{S}_k|^{-1/2} * EXP\left[(-.5)(\underline{X}_{ik}^T * \underline{S}_k^{-1} * \underline{X}_{ik})\right]}.$$

Tatsuoka's procedure (1971) incorporates the possibility of unequal variance-covariance matrices and prior probabilities which produce the posterior probabilities

$$P(j|\underline{x}_i) = \frac{PR(j) * P_j}{PR(I) * P_I + PR(II) * P_{II}} \quad j = 1,2 \text{ and } i = 1,2,\cdots N.$$

For computational purposes, the equation for $P(j|\underline{x}_i)$ is commonly presented as

$$P(j|\underline{x}_i) = \frac{EXP\left[(-.5)(D_j^2(\underline{x}_i) + LN|\underline{S}_j| - 2*LN(PR(j))\right]}{\sum_{k=I,II} EXP\left[(-.5)D_k^2(\underline{x}_i) + LN|\underline{S}_k| - 2*LN(PR(k))\right]}$$

where $D_j^2(\underline{x}_i) = \underline{X}_{ij}^T * \underline{S}_j^{-1} * \underline{X}_{ij}$. When the pooled variance-covariance matrix is to be used $D_j^2(\underline{x}_i) = \underline{X}_{ij}^T * \underline{S}_p^{-1} * \underline{X}_{ij}$ and the terms $LN|\underline{S}_k|$ are omitted. These formulas for $P(j|\underline{x}_i)$ are identical to those being used by SAS User's Guide: Statistics (1982) and can be shown to produce identical results as the method previously discussed from Marascuilo and Levin (1983).

In the development of prediction procedures, such as those previously presented, many researchers assume the use of an additive model, i.e., the lack of interaction effects. However, this is not always the case, and failing to include appropriate interaction variables will result in reduced prediction efficiency. The common use of additive models probably stems from the difficulty involved in determining which interaction terms are appropriate. Most studies will develop interaction variables by pair-wise multiplications of original variables, i.e., cross product terms such as $x_3 = x_1 * x_2$. The variable x_3 is then used to represent interaction between variables x_1 and x_2. The problem with this approach is that not all interaction effects can be

modeled this easily. It is conceivable that the interaction between x_1 and x_2 is really measured by $x_1 \cdot x_2^3$ or even something as complex as $exp\sqrt{x_1 \cdot x_2^3}$. When there are numerous predictors, there may also exist higher order interactions. In this case, the researcher could construct what would appear to be an endless number of interaction terms in an attempt to fit the nonadditive model. For example, when using 20 predictor variables, there may be $\binom{20}{2} = (20!)/[2!)\cdot(18!)] = 190$ interaction terms needed to model all possible simple interactions. Next, the researcher would consider all 1140 possible triple interactions and so forth. Obviously, this approach can quickly get out of hand. An additional complication with the above procedure is that all interaction effects are measured over the entire sample, when, in fact, they may only be significant in certain subgroups of the sample space.

There is a relatively new procedure which can automatically account for all interaction effects present in the model. This procedure is entitled Automatic Interaction Detector (AID) and was developed by Sonquist et al. (1971). The procedure maximizes the explanation of criterion variance through a sequence of splitting the original data space into several subgroups. Each split is based on a specific predictor variable which will maximize the between subgroup sum of squares and minimize the error or within subgroup sum of squares. Through this iterative splitting

procedure, the original sample is reduced to mutually exclusive subgroups which have little or no interaction of the remaining predictor variables, i.e., the prediction model within each subgroup is now additive. Empirical studies of the AID procedure, such as that by Karathanos (1975), have revealed the increased efficiency of AID over multiple regression when interaction effects are present.

Several other procedures have been used in modeling the persister/dropout problem; however, they are explanatory in nature and do not necessarily enhance prediction. For example, path analytic and linear structural relations (LISREL) models are useful for modeling causality and multicollinearity. But from a prediction standpoint, there is no improvement beyond a multiple regression model. Similarly, factor analysis is used to group predictor variables and to develop independent factors to be used in the regression or discriminant models. This process may enhance a deeper understanding of persister or dropout characteristics, but it will not improve the predictability of the model. In addition, researchers using these procedures are also faced with resolving the problem of linearity and additivity assumptions.

Summary

The survey of literature has revealed a set of predictor variables that have frequently been found to be

significant contributors in persister/dropout models. These variables include high school rank and GPA, SAT and ACT scores, and gender. In most studies race is either poorly coded or not included. Other fallacies in previous works are the lack of longitudinal studies (i.e., studies which cover more than a one year period) and the proper modeling of interaction effects. Of all the studies that were reviewed, only one presented the proportion of correct classifications which is needed to measure the efficiency of the model.

Previous attrition studies at the U.S. Air Force Academy have not proved to be beneficial in providing a procedure which will aid in the reduction of the large numbers of dropouts. These studies are also incomplete in that they only predict freshman attrition, and they do not include gender or race as predictor variables. Considering the tremendous taxpayer costs to educate and train each cadet, it is obvious that a more accurate and updated prediction model would be beneficial to determine which qualified candidates should be afforded an appointment.

The two group discriminant procedure has emerged from the literature review as the state of the art methodology for predicting persisters and dropouts. Examples of two group discriminant procedures are presented by Marascuilo and Levin (1983) using L_i, and by Tatsuoka (1971) using $D_j^2(\underline{x}_i)$. However, these procedures do not provide a reasonable approach for handling interaction terms when additivity

assumptions are violated. A relatively new procedure which sequentially splits the data space into mutually exclusive subgroups where interaction effects are minimized is AID. As an independent methodology, AID has produced results which surpass multiple regression for some nonadditive models. The literature does not indicate previous attempts to combine AID with discriminant analysis; however, such a combination appears to have the following advantages: (a) when there are numerous predictor variables the problem of investigating specific interaction terms is eliminated; (b) after splitting the original data into several subgroups that more closely meet additivity assumptions, the remaining predictor variables continue to be utilized by the discriminant model; (c) many researchers subgroup their data based on variables of interest and convenience that can impede the prediction process; whereas, when flexibility in subgrouping is feasible, AID provides a systematic splitting that will most likely enhance prediction.

CHAPTER III

PROCEDURES

Subjects

The data to be used for this study are from the cadet classes entering the United States Air Force Academy during the summers of 1979 and 1980, the graduating classes of 1983 and 1984 respectively. Although these classes are studied almost in their entirety, they are in reality, samples from the population of all cadets who have attended USAFA since its founding in 1955. These two classes are in close time proximity but differ in size. In addition, four years of attrition data are available for the class of 1983; however, only the first three years of attrition data are available for the class of 1984. Since less than 2% of the cadets drop out in their last year, this difference should not substantially effect the class prediction models but will probably have a slight impact on the cross-validation analysis. Although these differences in classes are undesirable, it was concluded that the differences should not overly distort the model development.

Subjects in either class who had been disenrolled for such atypical reasons as death or medical disqualification were deleted from the study. Also, students who were classified as stopouts, suspensions or turnbacks were not included. In all, 45 such subjects from the class of 1983 and 59 subjects from the class of 1984 were deleted from the study. Therefore, the number of cadets to be included in the study is 1463 for the class of 1983 and 1549 for the class of 1984. The deletion of subjects mentioned above is not anticipated to effect the overall outcome of the analysis; rather, it is an attempt to avoid obtaining confounding results.

Criterion Variable

This study utilized two predictor models in which each cadet was classified as either a persister or dropout. For multiple regression models, these two groups are binary coded. For example, persisters are coded 0 and dropouts are coded 1. However, in two group discriminant analyses, the groups must be identified, but coding is not required.

Predictor Variables

The SVIB variable information which Dempsey and Fast (1976) found to be very significant was not available. The absence of these variables might possibly limit the predictive efficiency of the models but will not distort the model

comparisons to be made. Predictor variables included in this study are:

a. Prior Academic Record (HSPAR) is a weighted composite variable based on academic rank, GPA, high school size and membership in an honor society. The weights are determined by the Air Force Academy Registrar's Office;

b. College Board Aptitude and Achievement Scores to include verbal (ATPVERB) and math aptitude (ATPMATH);

c. American College Test scores for English (ACTENGL) Social Studies (ACTSS), Mathematics (ACTMATH) and Natural Science (ACTNS);

d. Athletic Activity index (ATXCND) based upon athletic participation in high school or community athletic programs;

e. Physical Aptitude Examination (PAECND) computed from a rigorous physical test;

f. Non-Athletic Activity Index (NATXCND) is based on nonathletic high school and community activities such as president of a class, major role in a play and state or regional speech or debate;

g. High school size (HSCLS);

h. High school rank (HSRANK), converted to a standardized score ranging from 200 to 800 (μ = 500, σ = 100);

i. Academic composite score (ACACMP), computed by USAFA Registrar based on ACT and SAT scores;

j. Gender (coded 0 for males and 1 for females);

k. Recruited athlete (ATSPCND), 0 = no, 1 = yes;

l. Prior college, 0 = no, 1 = yes, referred to as CLGATTN;

m. Attended a military prep school (PREPATN), 0 = no, 1 = yes;

n. Race (binary coded for Whites (R1), Blacks (R2), Hispanics (R3), Orientals (R4), and other). This results in four binary variables for group membership by race;

o. Parent's prior academy (ACADPNN) where 0 = no, 1 = yes;

p. Parent's prior military (MILSTSP) where 0 = no, 1 = yes.

Major Research Objectives

The primary emphasis of this study is to develop a new prediction procedure which combines MAID and two group discriminant analysis. Comparing this new procedure with the usual two group discriminant analysis provide a means of evaluating the effectivenss of incorporating MAID in a prediction model. With this in mind, the major research objectives for this study are stated as follows:

Objective 1. Using a two group discriminant procedure on two large samples of admissions data determine the hit rate and the amount of shrinkage obtained when predicting persisters and dropouts.

Objective 2. Develop a new prediction procedure which combines a modified Automatic Interaction Detector (MAID) with two group discriminant analysis.

Objective 3. Using the procedure developed in Objective 2 and the two large samples of college admissions data, determine the hit rate and amount of shrinkage obtained in predicting persisters and dropouts.

Objective 4. Determine the unique contribution of incorporating MAID in a persister/dropout model.

Procedures Used to Answer Major Research Objectives

Procedures for Objective 1. The two group discriminant procedure found in the SAS User's Guide: Statistics (1982) will be applied to the data obtained from the Academy. This procedure consists of determining two posterior probabilities for each subject, $P(j|\underline{x}_i)$ for j=1 or 2. $P(j|\underline{x}_i)$ is defined as the probability that subject i is in group j given that he or she obtained the variable vector of scores $\underline{x}_i$, for the p predictors. This probability is mathematically represented as:

$$P(j|\underline{x}_i)=EXP\left[(-.5)(D_j^2(\underline{x}_i)+LN|\underline{S}_j| -2*LN(PR(j)))\right] / \underset{k}{SUM}$$
$$\left[EXP(-.5)(D_k^2(\underline{x}_i)+LN|\underline{S}_k|-2*LN(PR(k)))\right].$$

The following definitions will clarify any notation problems:

j is either 1 for group I or 2 for group II

$P(j|\underline{x}_i)$ is the posterior probability that subject i is in group j given a set of scores, $\underline{x}_i = (x_{i1}, x_{i2}, ---, x_{ip})$ for subject i on the p predictor variables.

EXP is the exponential function.

$D_j^2(\underline{x}_i)$ is the squared generalized distance between a vector of scores $\underline{x}_i$ for subject i on a set of p predictors and the group j centroid represented by $(\bar{x}_{j1}, \bar{x}_{j2}, \ldots, \bar{x}_{jp})$.

Thus, $D_j^2(\underline{x}_i) = (x_{i1} - \bar{x}_{j1}, x_{i2} - \bar{x}_{j2}, \ldots, x_{ip} - x_{jp})^T * \underline{S}_j^{-1} * (x_{i1} - \bar{x}_{j1}, x_{i2} - \bar{x}_{j2}, \ldots, x_{ip} - \bar{x}_{jp})$.

This is a distance measuring procedure developed by Mahalanobis (Marasciulo and Levin, 1983).

$\underset{k}{\text{Sum}}$ refers to the sum over all possible values of k which for two group discriminant analysis is k = 1, 2 for groups I and II.

$D_k^2(\underline{x}_i)$ refers to the squared generalized distance of subject i's vector of scores $\underline{x}_i = (x_{i1}, x_{i2}, ---, x_{ip})$ from the kth group centroid $(\bar{x}_{k1}, \bar{x}_{k2}, \ldots, \bar{x}_{kp})$.

LN is the natural logarithm function.

$|\underline{S}_k|$ is the determinant of the variance-covariance matrix for group k.

PR(k) is the prior probability that subject i is from group k.

The posterior probability, $P(j|\underline{x}_i)$ is computed for each group given subject i. Subject i is then classified into the group j which provides the largest posterior probability. The terms $LN|\underline{S}_j|$ and $LN|\underline{S}_k|$ are only used when it has been determined by way of a chi-squared test that $\underline{C}_I \neq \underline{C}_{II}$ (where $\underline{C}_I$ is the population variance-covariance matrix for group I). If the pooled variance-covariance matrix is used then $D_j^2(\underline{x}_i)$ uses $\underline{S}_P^{-1}$ instead of $\underline{S}_j^{-1}$. The terms LN(PR(j)) and LN(PR(k)) will vary based on what prior probabilities the researcher desires to use. Since both samples for this study are very large, the prior probabilities will be defined as $PR(I) = N_I/(N_I+N_{II})$ and $PR(II)=N_{II}/(N_I+N_{II})$, where N_I and N_{II} are the group I and group II sample sizes. The predicted and actual persisters and dropouts will be displayed in a contingency table along with the overall hit rate. Since the class of 1983 experienced an attrition rate of 36.6%, a prediction model hit rate must exceed the base rate of .634 for the model to be considered effective. In addition, the prediction model for the class of 1983 will be cross-validated on the class of 1984. The difference in hit rates for these two samples is referred to as shrinkage and will be presented as a measure of generalizability of the procedure.

Procedures for Objective 2. The combination of two group discriminant analysis and MAID into a new procedure can be performed as a two-phase process. First, conduct the MAID procedure to split the data into mutually exclusive

subgroups which will more closely meet additivity assumptions. The second step requires performing two group discriminant analysis within each subgroup.

The modified AID procedure is initiated with the search of a predictor variable which will be used to split the original data set in such a way as to reduce the error variance. This is accomplished when the splitting process results in maximizing the sum of squares between groups (SSB) and in decreasing the sum of squares within groups (SSW). The sum of squares total (SST) will remain the same throughout, and, therefore, RSQ = SSB/SST will increase as should predictive efficiency. In order to maximize SSB and minimize SSW during each split, the predictor variable, x_s, with the highest validity coefficient, r_{yx_s}, is chosen as the variable to split the data set. Once the initial split has been accomplished, the priority of data subgroups to be subsequently split is based on the subgroup with the maximum total sum of squares i.e. $TSS_j = \sum_{i=1}^{n_j} (Y_{ij} - \hat{Y}_{ij})^2$. This process continues until one of the following termination criteria is met: (a) the maximum number of desired splits is attained; (b) the percent of total sums of squares in each existing subgroup is less than the minimum percentage desired for further splitting; or (c), the sample size in each existing subgroup is too small to justify further splitting. Summarizing MAID in a step-by-step procedure results in the following:

Step 1. Given a data set where $N \geq 1000$, determine all validity coefficients, i.e., simple correlation coefficients of the dependent variable and each of the predictor variables. Choose the variable which has the largest validity coefficient and split the original data set into two subgroups. If the selected predictor variable, x_s, is a dichotomous variable, the initial split is simply into two subgroups; one where $x_s = 0$ and the other where $x_s = 1$. If the predictor variable selected for the split is continuous, the researcher must determine a cutoff point for that variable which will result in maximizing efficiency. A previously discussed cutoff determination procedure for two group discriminant analysis will be used, i.e.,

$$c = \frac{N_I \frac{\bar{X}_{sI}}{S_{X_{sI}}} + N_{II} \frac{\bar{X}_{sII}}{S_{X_{sII}}}}{\frac{N_I}{S_{X_{sI}}} + \frac{N_{II}}{S_{X_{sII}}}}$$

Step 2. Determine which unsplit subgroup j has the largest TSS_j. Use this subgroup for the next split.

Step 3. Given the subgroup identified in Step 2, determine the predictor variable, x_s', with the largest validity coefficient.

Step 4. Determine the proper cutoff criterion for x_s' (as described in step 1) for splitting the subgroup.

Step 5. Examine the current set of data subgroups to determine if any of the preselected termination criteria have been met. If termination criteria are met, discontinue the splitting process; if not, return to Step 2.

After the MAID procedure is completed, conduct a two group discriminant analysis, as presented in Objective 1, for each final subgroup. The overall hit rate is obtained by summing the number of correctly classified subjects over all mutually exclusive subgroups and dividing by the sample size.

To perform a cross-validation of this new procedure, first force the same splitting procedure found with the initial sample on the second sample. Then use the discriminant model developed for each final subgroup in the first sample to predict subjects in the corresponding second sample subgroups. As previously discussed, the hit rate is obtained by summing correctly classified subjects over each final subgroup and dividing by sample size. The amount of shrinkage is then computed as the difference in the first sample hit rate and the second sample hit rate.

Procedures for Objective 3. This objective will be accomplished in the following manner: (a) perform the splitting process described in Objective 2 on the class of 1983 until $N_j \leq (.10)N_T$ or $TSS_j \leq (.10)TSS_T$ for all unsplit subgroups or until the total number of splits is equal to 10; (b) force the same splitting procedures on the class of 1984; (c) conduct a two group discriminant analysis for each final subgroup in the class of 1983 and cross-validate the model on each corresponding subgroup in the class of 1984; (d) determine the hit rate for each class by summing correctly classified subjects over each final subgroup and divide by

the sample size; (e) determine the amount of shrinkage by subtracting the hit rate for the class of 1984 from the hit rate for the class of 1983.

Procedures for Objective 4. The unique contribution of MAID is defined as the amount of predictability contributed by MAID above and beyond that of the two group discriminant model. This unique contribution is estimated by the difference in hit rates of the discriminant model and the combined model. In addition, the difference in shrinkage for these two models will also be reported to evaluate the possible usefulness of the new procedure.

CHAPTER IV

FINDINGS AND INTERPRETATIONS

This chapter is designed to present and interpret the results obtained after implementing the procedures for each objective discussed in the previous chapter. All of the computer work was performed on an IBM 370 computer system utilizing SAS: 1982 Edition software.

Findings

Results for Objective 1. A two group discriminant procedure was applied to two large samples of admissions data from the Air Force Academy. The descriptive statistics for each variable in both samples are found in Table 5. The Class of 1983 was used for the initial analysis and the resulting discriminant function was used to predict success and failure for subjects in the Class of 1984. The results for the Class of 1983 are summarized in Table 6 and as follows:

1. χ^2 (test for homogeneity of variance) = 1432.599.
2. Degrees of freedom = 276.
3. χ^2 is significant at the .01 level.
4. Hit rate = 843/1463 = .5762.

The results obtained from the cross-validation on the Class of 1984 are presented in Table 7. The hit rate for the Class of 1984 is 784/1549 = .5061. As previously defined, the amount of shrinkage in the model is the difference in the two hit rates which is .5762 - .5061 = .0701.

Table 5

Descriptive Statistics for Variables in Both Data Samples

	Class of 1983 (N=1463)			Class of 1984 (N=1549)		
ariable	Mean	Std Dev.	Sum	Mean	Std Dev.	Sum
DROPOUT	.374	.484	547	.352	.478	546
MILSTSP	.176	.381	257	.221	.415	342
ACACMP	3106.456	293.945	4544584	3094.855	306.843	4793931
CLGATTN	.098	.297	143	.076	.265	118
PREPATN	.165	.371	241	.154	.361	238
ACADPNN	.012	.107	17	.022	.147	34
HSPAR	612.822	96.428	896558	616.660	94.000	955206
ACTSS	26.119	3.572	38212	26.204	3.476	40590
ACTNS	28.707	3.007	41998	28.804	2.883	44618
ACTMATH	28.510	2.928	41710	28.624	2.866	44338
ACTENGL	22.586	2.959	33043	22.850	2.989	35395
R1	.848	.359	1240	.837	.370	1296
R2	.073	.260	107	.073	.260	113
R3	.042	.200	61	.061	.240	95
R4	.031	.175	46	.023	.151	36
GENDER	.115	.319	168	.138	.345	214
ATPVERB	54.726	7.127	80064	54.086	7.134	83779
ATPMATH	63.774	6.822	93302	63.037	6.715	97645
PAECND	540.211	84.078	790328	528.542	78.121	818712
ATXCND	537.841	107.157	786861	535.891	107.165	830095
NATXCNP	530.178	103.532	775651	543.321	98.506	841605
ATSPCND	.169	.375	247	.195	.396	302
HSCLS	367.118	185.125	537093	373.949	206.812	579247
HSRANK	20.366	25.642	29795	20.708	24.921	32077

Results for Objective 2. The new procedure which combines discriminant analysis and MAID will be referred to as MAIDDA and can be applied to any data set which has a dichotomous dependent variable, numerous categorical and/or continuous independent variables, and a sample size greater

than 1000. The proportions for each group of the dependent variable should normally not be outside the .20 and .80

Table 6

Class of 1983 Prediction Results

		ACTUAL	
	Dropouts	Persisters	TOTAL
Dropouts	499	572	1071
PREDICTED			
Persisters	48	344	392
TOTAL	547	916	1463
Percent correctly classified	91.22%	37.5%	57.62%

Table 7

Class of 1984 Prediction Results

		ACTUAL	
	Dropouts	Persisters	TOTAL
Dropouts	422	641	1063
PREDICTED			
Persisters	124	362	486
TOTAL	546	1003	1549
Percent correctly classified	77.29%	36.09%	50.61%

range and any categorical independent variables should be binary coded. The step-by-step process for MAIDDA is as follows:

Step 1. For the original sample, determine simple correlation coefficients of the dependent variable with each independent variable; these values are also referred to as the validity coefficients. Select the variable, X_S, which produces the largest validity coefficient as the variable to be used to split the data set. In this case, largest refers to the maximum of the validity coefficients without regard to positive or negative signs.

Step 2. The splitting process varies as to whether the independent variable is binary coded or continuous. If the variable is binary coded, the split is simply into one subset where $X_S=0$ and another where $X_S=1$. If the variable is continuous, determine a cutoff point, c, as follows:

$$C = \frac{N_I \frac{\bar{X}_{SI}}{S_{X_{SI}}} + N_{II} \frac{\bar{X}_{SII}}{S_{X_{SII}}}}{\frac{N_I}{S_{X_{SI}}} + \frac{N_{II}}{S_{X_{SII}}}}$$

Then split the data set into one subgroup where $X_S < C$ and another where $X_S \geq C$. If a cross-validation is to be performed, force the same splits on the second data sample.

Step 3. Record the sequence of data splits in an upside down tree fashion. In other words, if node 1 represents the original sample, then Figure 3 displays the results of splitting the original sample into subgroup 2 where $X_S = 0$ and subgroup 3 where $X_S = 1$.

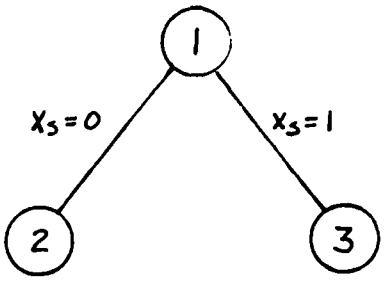


Figure 3. Example of a Tree Diagram for Data Splits

Step 4. For each unsplit subgroup j in the initial sample tree diagram, compute the total sum of squares, $TSS_j = \sum_{i=1}^{n_j} Y_i^2 - \left(\sum_{i=1}^{n_j} Y_i\right)^2 / n_j$.

Step 5. Compare the TSS_j for all unsplit subgroups and identify the subgroup which has the largest TSS_j. If $TSS_j \leq (.10)(TSS_1)$, or $N_j \leq (.10)N_1$ or the total number of splits is 10 then go to Step 7, otherwise continue. These termination criteria were determined in an attempt to reduce the possibility of an overfit condition, i.e. a low N_j/P ratio.

Step 6. Given the subgroup j found in Step 5, determine all simple correlation coefficients between the dependent variable and each predictor variable within that subgroup for the original sample. Select the predictor which produces the largest validity coefficient, X_S', as the new splitting variable for subgroup j. Return to Step 2.

Step 7. Perform a two group discriminant analysis for each final subgroup in the orginal sample, i.e., for each unsplit subgroup. Use each subgroup's discriminant function

to cross-validate the model on the corresponding subgroup in the second sample.

Step 8. In order to determine the system hit rate, sum up the number of subjects correctly classified in each sample's final subgroups and divide by the corresponding sample size. The amount of shrinkage in the model is then determined by subtracting the second sample hit rate from that of the original sample.

Results for Objective 3. The MAIDDA procedure described above was applied to the same two samples used in Objective 1. The results of Step 1 for the Class of 1983 data are shown in Table 8.

Table 8

Simple Correlation Coefficients for the Class of 1983 Data

VARIABLE	VALIDITY COEFFICIENT	VARIABLE	VALIDITY COEFFICIENT
CLGATTN	.0311	R3	.0084
PREPATN	-.0385	R4	-.0340
MILSTSP	-.3010*	R4	-.0340
ACADPNN	-.0574	GENDER	.0407
HSPAR	-.1698	ATPVERB	-.1067
ACACMP	-.2129	ATPMATH	-.1209
ACTSS	-.1029	PAECNG	.0497
ACTNS	-.1032	ATXCND	.0344
ACTMATH	-.0796	NATXCND	-.0128
ACTENGL	-.0623	ATSPCND	.0515
R1	-.0182	HSCLS	-.0198
R2	.0488	HSRANK	.0864

Since MILSTSP, the binary variable representing military parents and non-military parents, had the largest

validity coefficient for the Class of 1983 sample, the first split appears as shown in Figure 4.

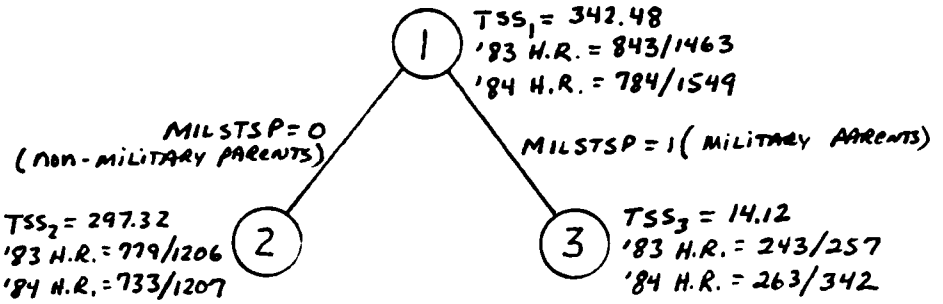


Figure 4. Tree Diagram after the First Split

The sum of squares total for subgroups 2 and 3 are TSS_2 = 297.324 and TSS_3 = 14.12; therefore, the second split will be on subgroup 2. Since ACACMP has the largest validity coefficient in subgroup 2, the cutoff point for making the split is $C = \frac{674(3154.5/272.84) + 532(3026.8/310.5)}{(674/272.8) + (532/310.5)} = 3101.$

The second split appears as shown in Figure 5.

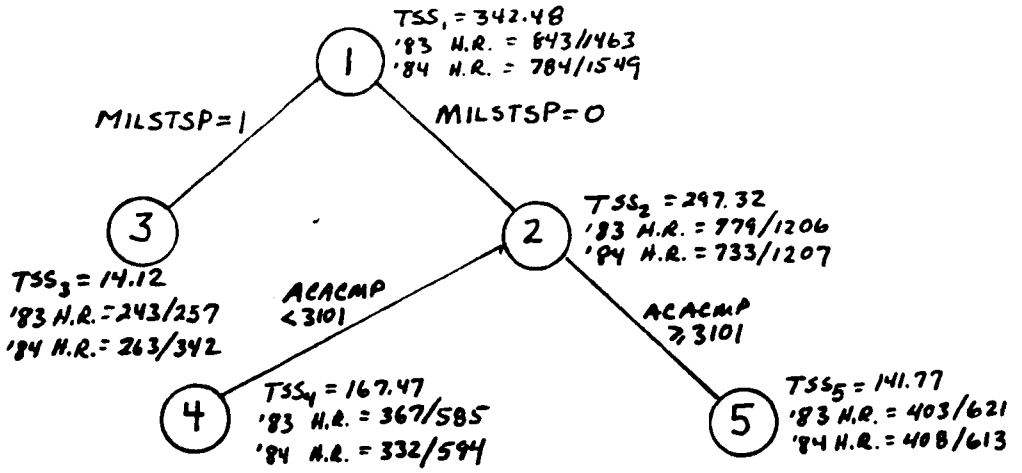


Figure 5. Tree Diagram after Two Splits

The sum of squares total for subgroups 4 and 5 are 167.47 and 141.77 respectively. The next split will be on subgroup 4 where the variable ACACMP again has the largest validity coefficient. The new cutoff point will be

$$C = \frac{272(2887.6/156.1) + 313(2816.8/194.7)}{(272/156.1) + (313/194.7)} \doteq 2853.$$

The third split produced a tree diagram as shown in Figure 6.

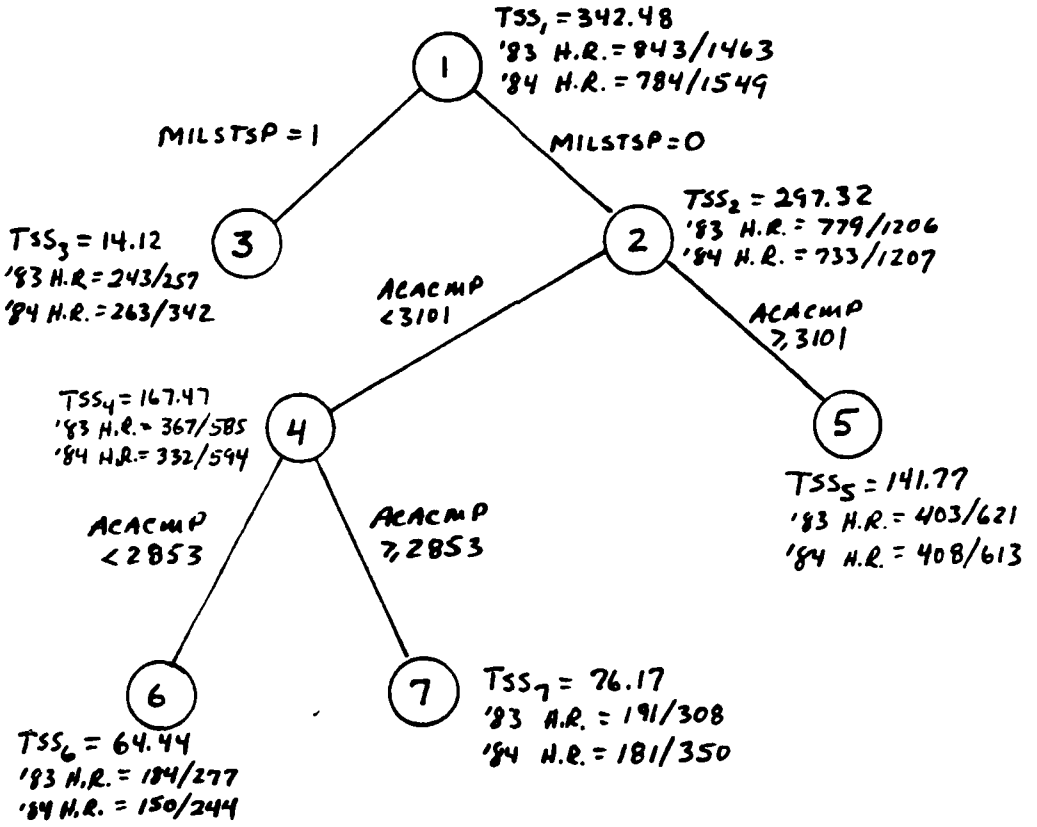


Figure 6. Tree Diagram after Three Splits

The computed sum of squares total for subgroups 6 and 7 are $TSS_6 = 64.44$ and $TSS_7 = 76.17$. Comparing TSSj for subgroups 3, 5, 6 and 7 reveals a maximum at $TSS_5 = 141.77$.

The fourth split takes place in subgroup 5 which has the largest validity coefficient for the variable GENDER. The splitting of subgroup 5 will be into subgroup 8 and 9.

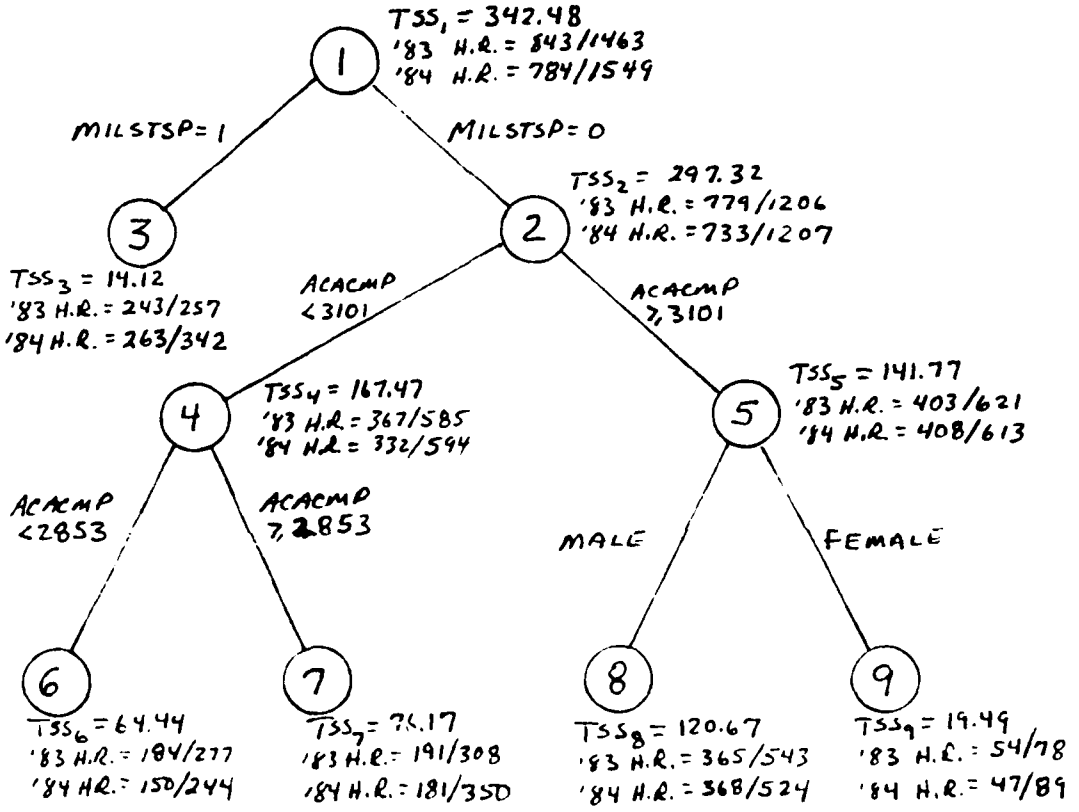


Figure 7. Tree Diagram after Four Splits

The sum of squares total for subgroups 8 and 9 are TSS_8 = 120.67 and TSS_9 = 19.49. The maximum TSSj for subgoups 3, 6, 7, 8 and 9 is TSS_8 = 120.167. The largest validity coefficient for subgroup 8 is .0742 for variable ACTMATH. The fifth split is then on subgroup 8 using ACTMATH < 29.7 and ACTMATH ≥ 29.7. The results after five splits are seen in Figure 8.

The fourth split takes place in subgroup 5 which has the largest validity coefficient for the variable GENDER. The splitting of subgroup 5 will be into subgroup 8 and 9.

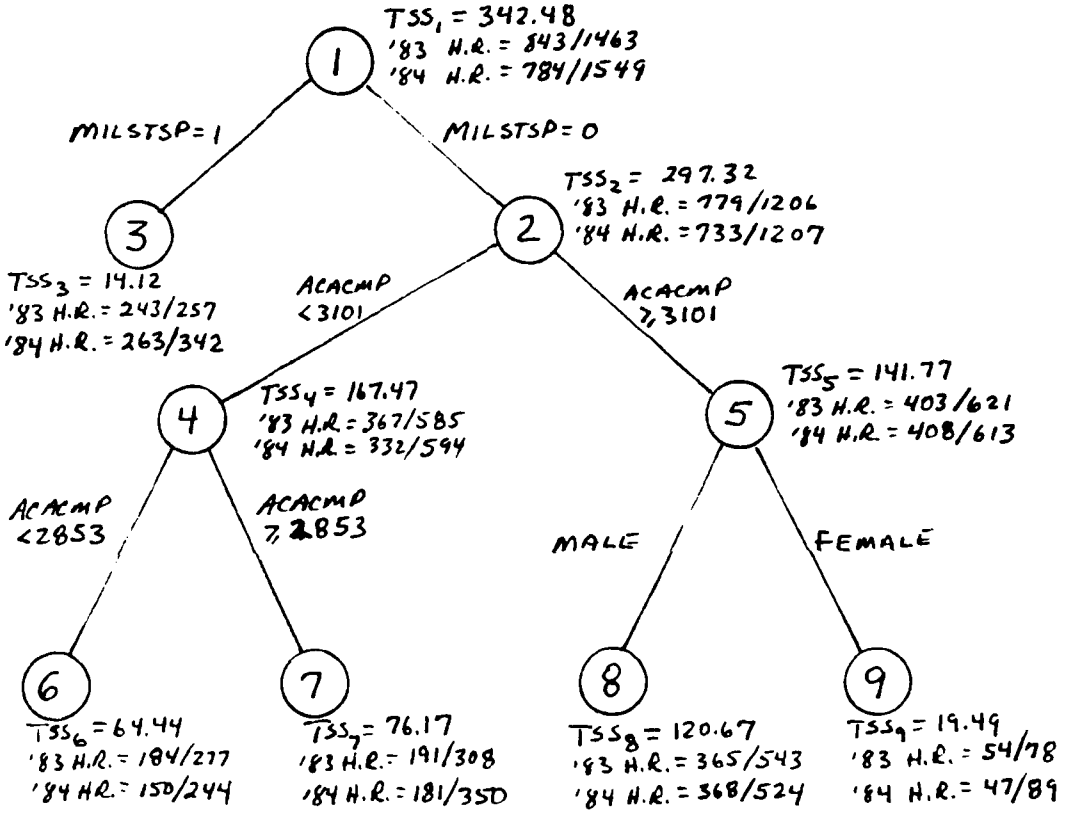


Figure 7. Tree Diagram after Four Splits

The sum of squares total for subgroups 8 and 9 are TSS_8 = 120.67 and TSS_9 = 19.49. The maximum TSSj for subgoups 3, 6, 7, 8 and 9 is TSS_8 = 120.167. The largest validity coefficient for subgroup 8 is .0742 for variable ACTMATH. The fifth split is then on subgroup 8 using ACTMATH < 29.7 and ACTMATH ≥ .29.7. The results after five splits are seen in Figure 8.

The fourth split takes place in subgroup 5 which has the largest validity coefficient for the variable GENDER. The splitting of subgroup 5 will be into subgroup 8 and 9.

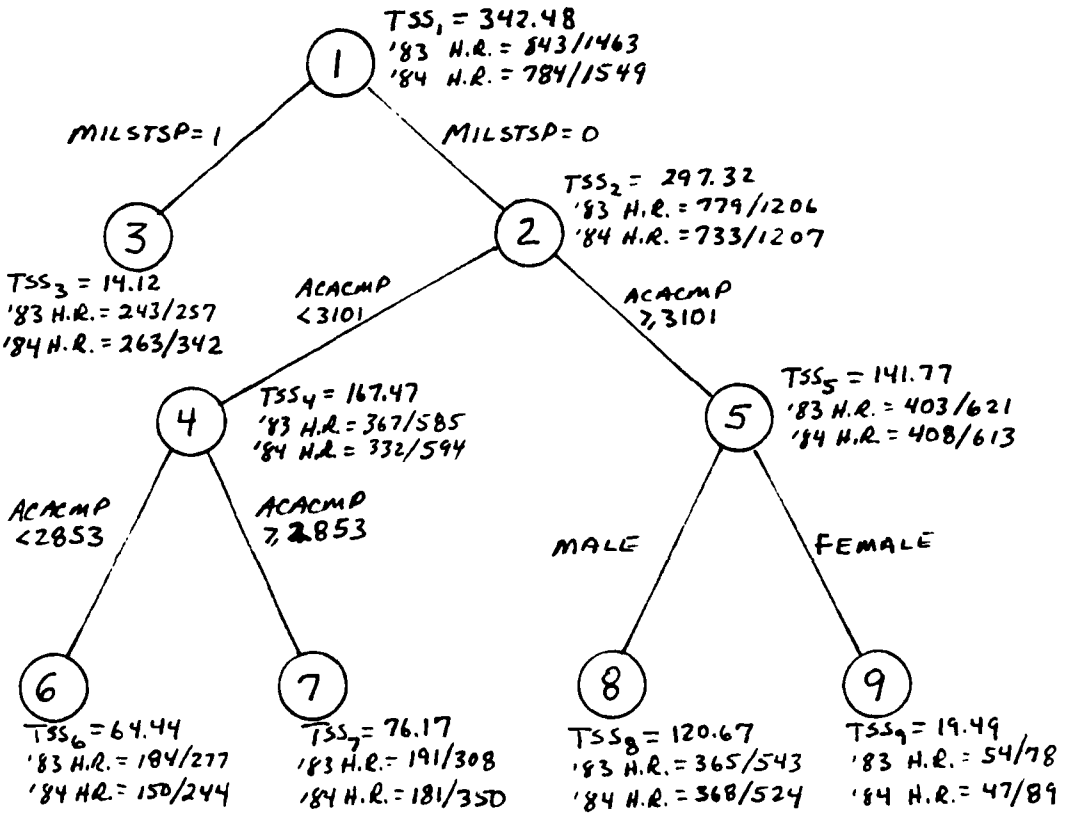


Figure 7. Tree Diagram after Four Splits

The sum of squares total for subgroups 8 and 9 are TSS_8 = 120.67 and TSS_9 = 19.49. The maximum TSSj for subgoups 3, 6, 7, 8 and 9 is TSS_8 = 120.167. The largest validity coefficient for subgroup 8 is .0742 for variable ACTMATH. The fifth split is then on subgroup 8 using ACTMATH < 29.7 and ACTMATH ≥ .29.7. The results after five splits are seen in Figure 8.

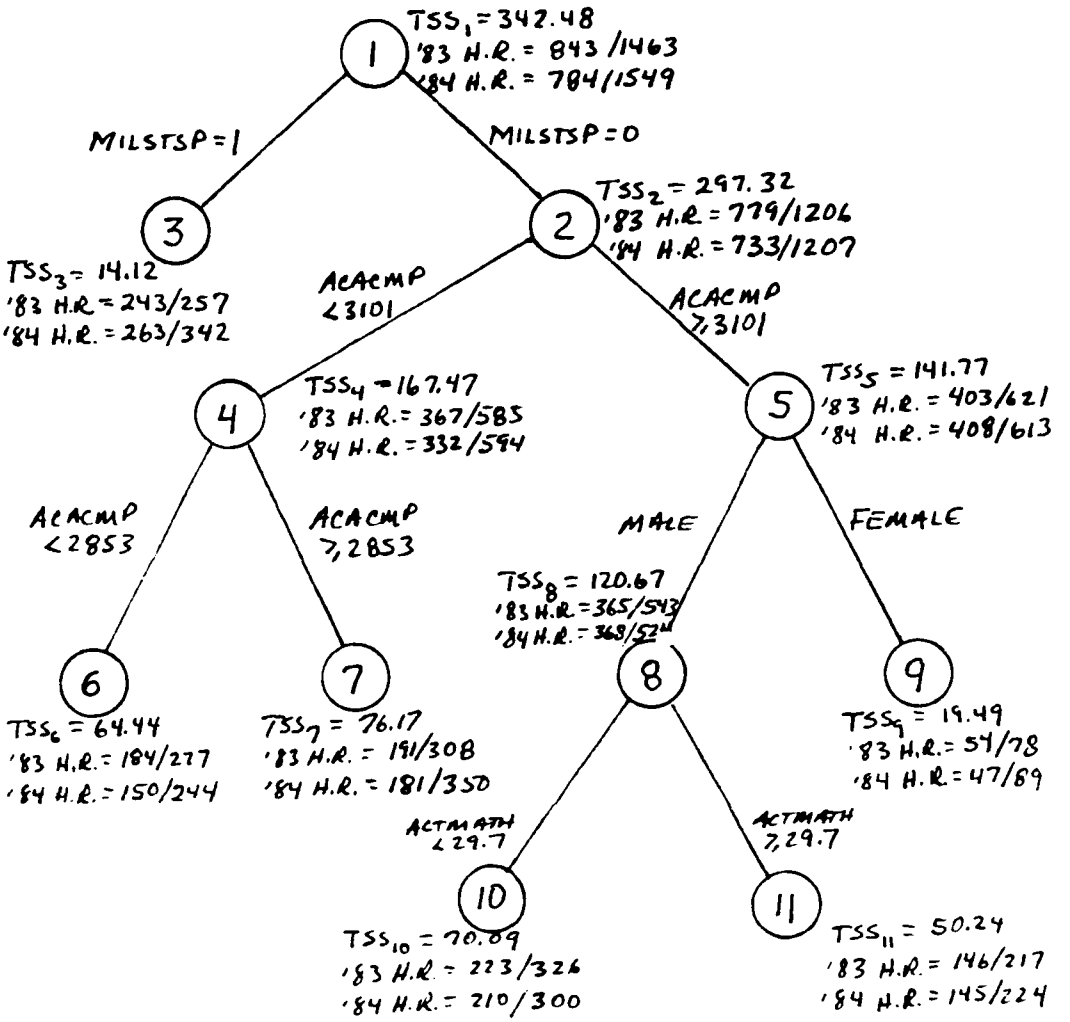


Figure 8. Tree Diagram after Five Splits

Total sums of squares for these two new subgroups are TSS_{10} = 70.09 and TSS_{11} = 50.24. The maximum TSS_j for extreme nodes 3, 6, 7, 9, 10 and 11 is TSS_7 = 76.17. Subgroup 7 will be split on the variable PAECND (Physical Aptitude Examination) which had a validity coefficient of .1000. Figure 9 depicts the corresponding tree diagram.

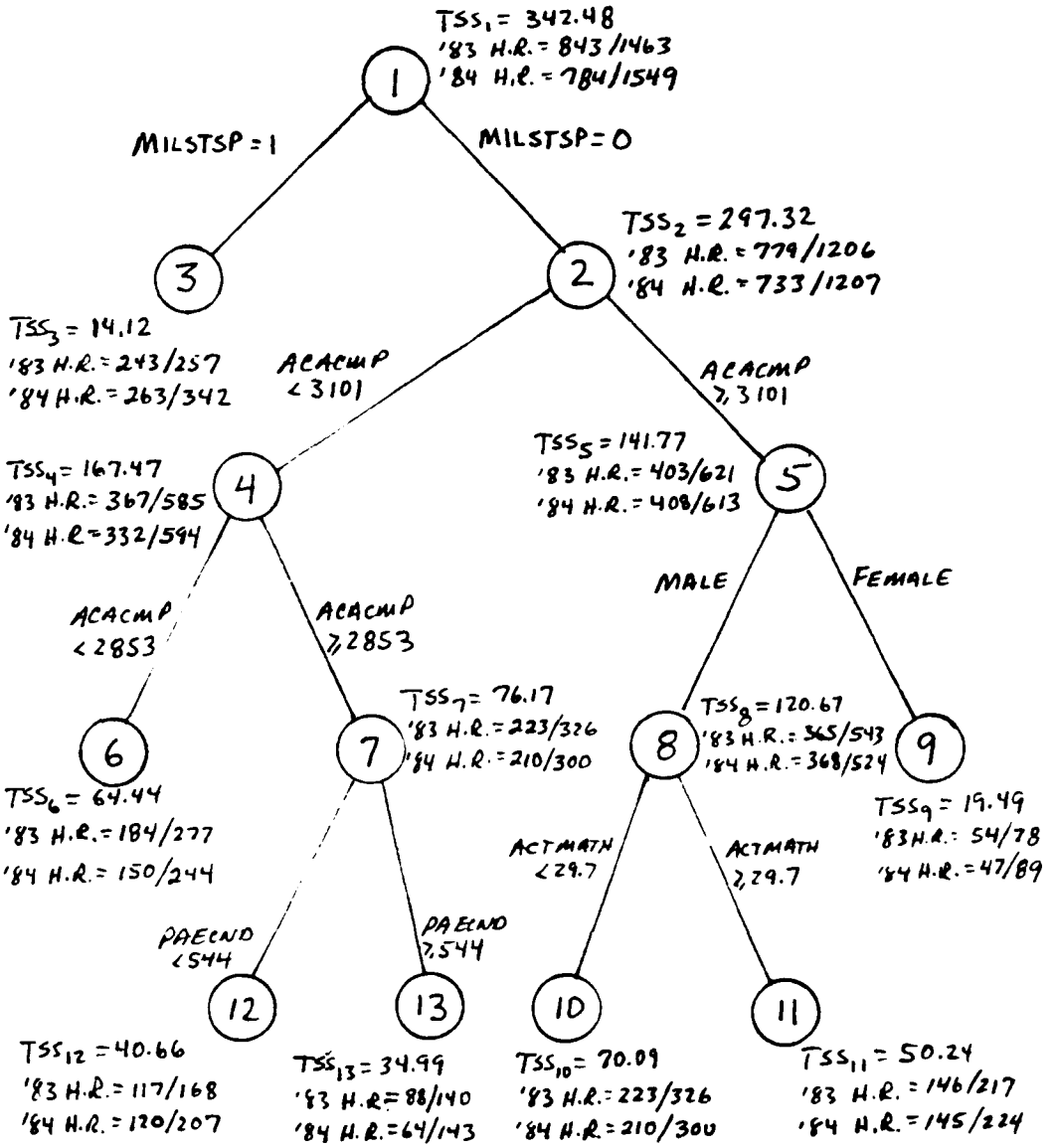


Figure 9. Tree Diagram after Six Splits

The seventh split is on subgroup 10 using ACTNS as the splitting variable, see Figure 10.

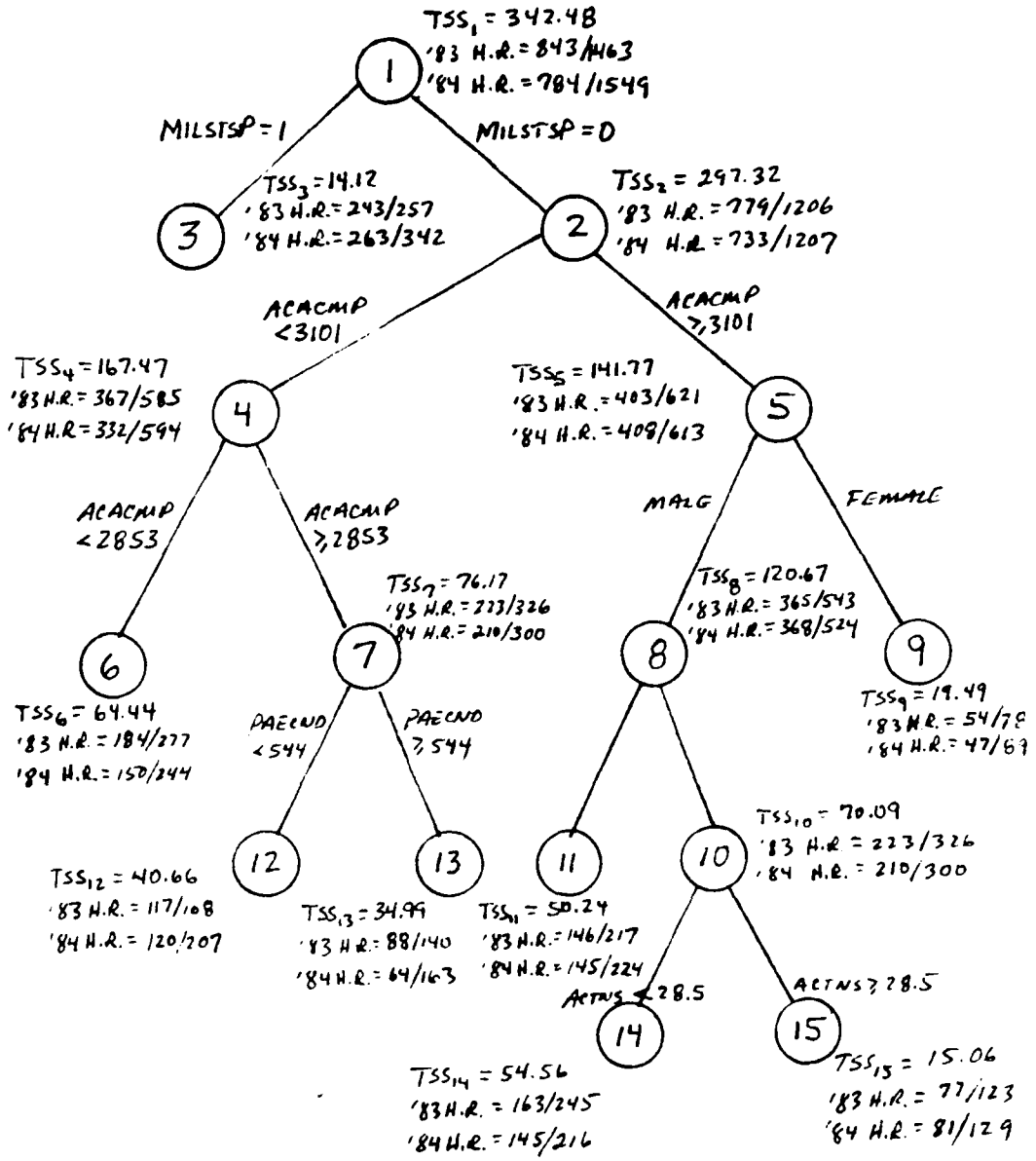


Figure 10. Tree Diagram after Seven Splits

Examining the current set of unsplit subgroups, (3, 6, 9, 11, 12, 13, 14, 15), the largest TSS_j occurs for Subgroup 6 which will be split on variable R1 (whites and non-whites). The data structure after the eighth split is shown in

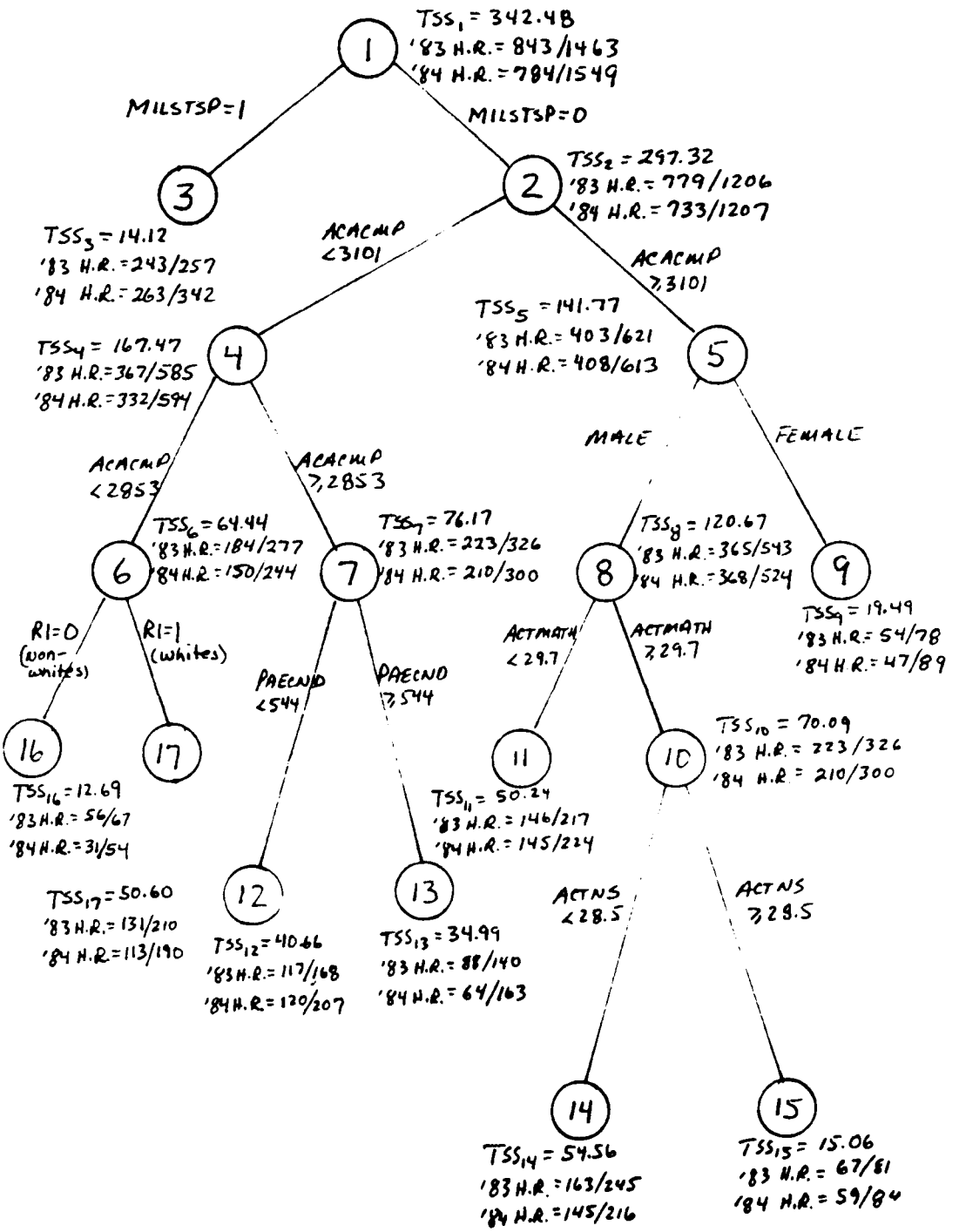


Figure 11. Tree Diagram after Eight Splits

Figure 11. The ninth split occurs on subgroup 14 where HSCL is the splitting variable. The cutoff point for HSCLS is 375.5 (See Figure 12). Since our termination criteria is $N_j \leq (.10)(N_1)$ or $TSS_j \leq (.10)(TSS_1)$ or number of splits equal 10, the next split is the last. The largest TSS_j for the current set of unsplit subgroups is $TSS_{17} = 50.6$ where HSPAR has the largest validity coefficient. The tenth and final split is presented in Figure 13.

The two sets of sample data have now each been split into eleven mutually exclusive subgroups which should be relatively free of predictor variable interaction. It is postulated that the prediction model within each subgroup now meets the additivity assumption, and the overall hit rate is increased from that of the discriminant analysis in Objective 1. The combined subgroup prediction tables for each class are presented in Tables 9 and 10. See Appendix A for individual subgroup prediction results.

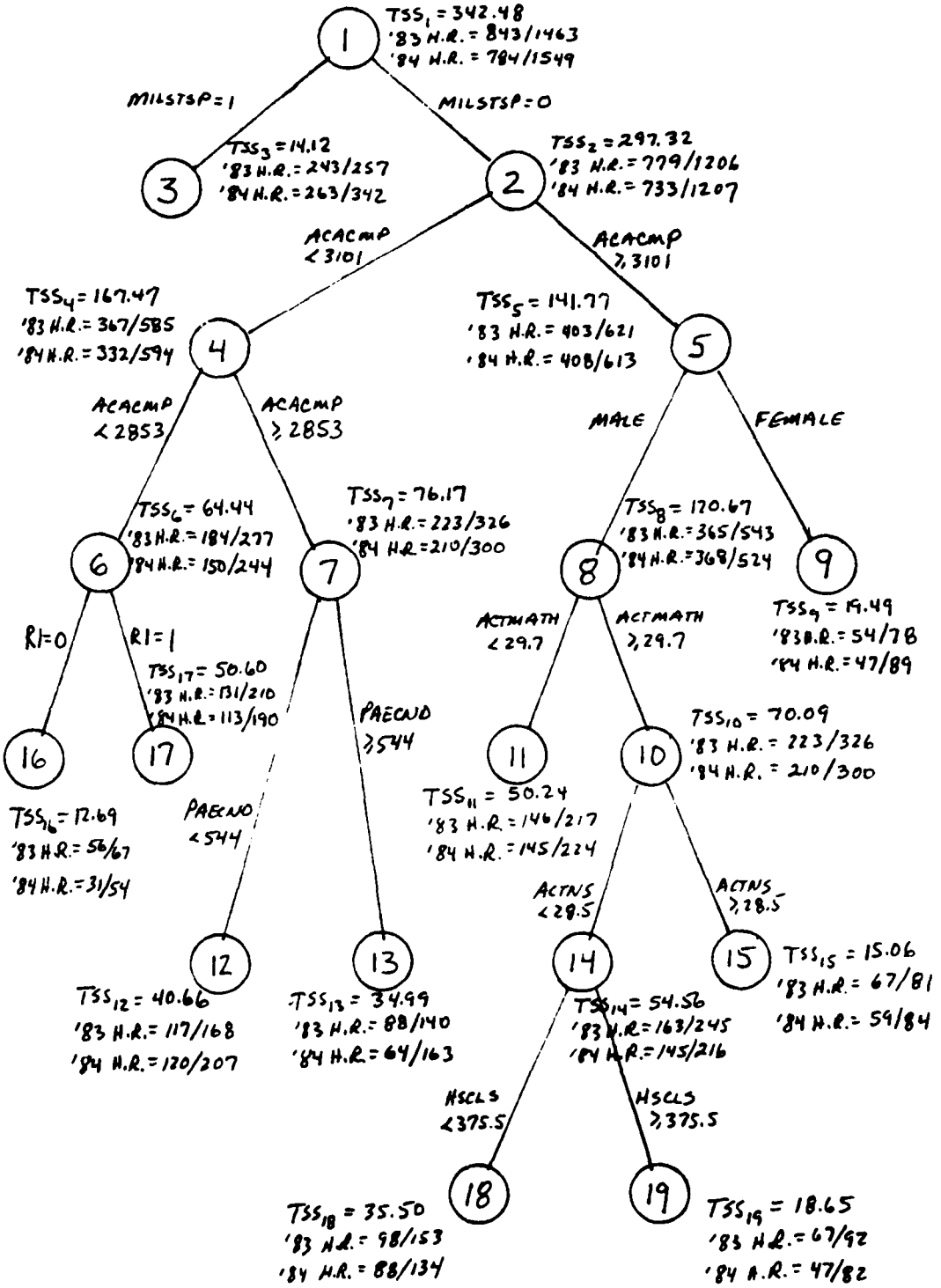


Figure 12 Tree Diagram after Nine Splits

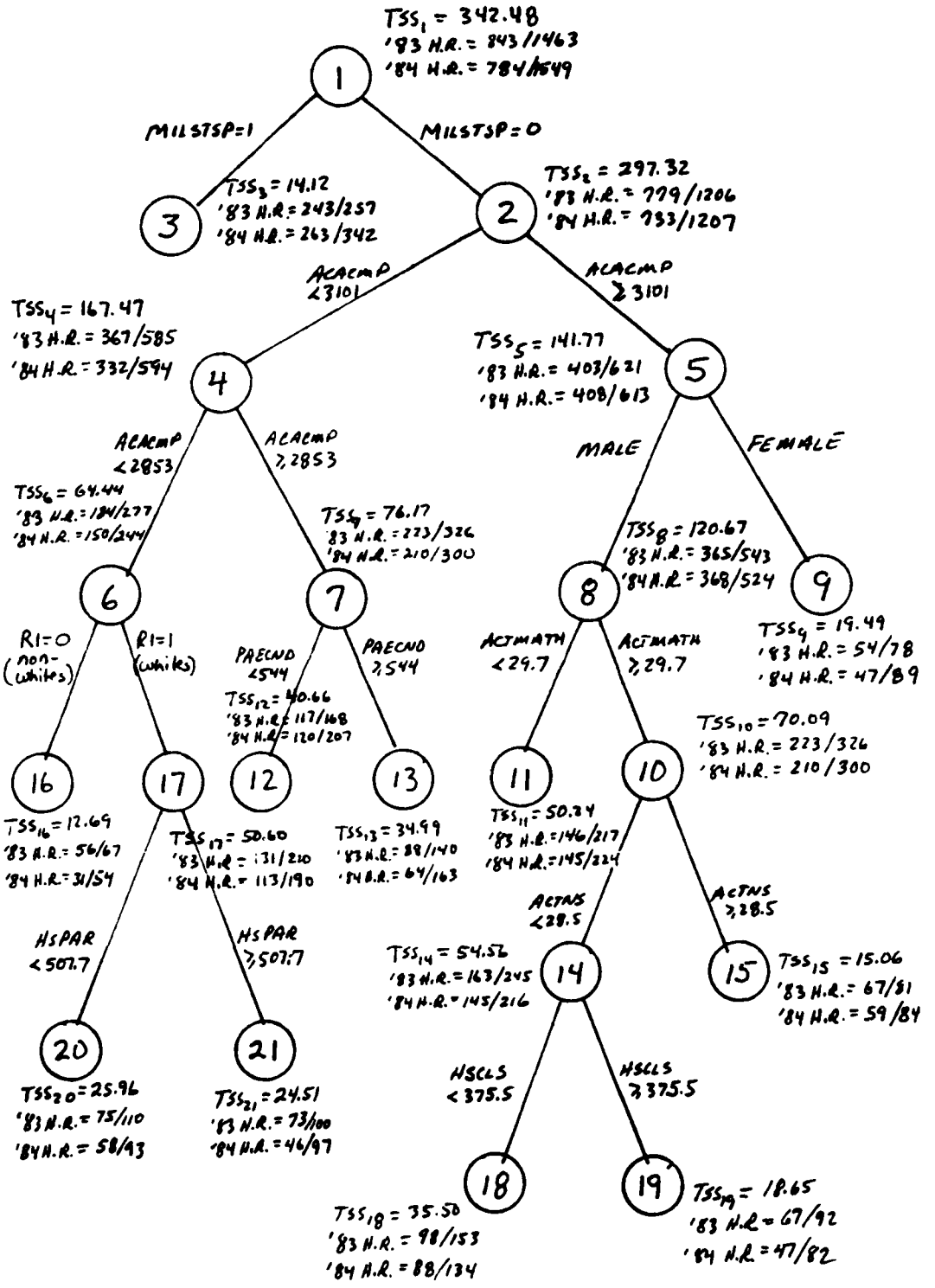


Figure 13 Tree Diagram after Ten Splits

Table 9

Prediction Results of Combined Subgroups 3, 9, 11, 12. 13, 15, 16, 18, 19, 20, and 21.

Class of 1983

	ACTUAL		
	Dropouts	Persisters	TOTAL
PREDICTED			
Dropouts	304	136	440
Persisters	243	780	1023
TOTAL	547	916	1463
Percent correctly classified	55.58%	85.15%	74.09%

Class of 1984

	ACTUAL		
	Dropouts	Persisters	TOTAL
PREDICTED			
Dropouts	202	235	437
Persisters	344	768	1112
TOTAL	546	1003	1549
Percent correctly classified	37.00%	76.57%	62.62%

Table 10

Hit Rate and Shrinkage Summary for Discriminant Analysis (Split # = 0) and MAIDDA (Splits 1 - 10)

Split #	'83 H.R.	'84 H.R.	Shrinkage
0	843/1463 = .5762	784/1549 = .5061	.0701
1	1022/1463 = .6986	996/1549 = .6430	.0556
2	1013/1463 = .6924	1003/1549 = .6475	.0449
3	1021/1463 = .6979	1002/1549 = .6469	.0510
4	1037/1463 = .7088	1009/1549 = .6514	.0574
5	1041/1463 = .7116	996/1549 = .6430	.0686
6	1055/1463 = .7211	1001/1549 = .6462	.0749
7	1062/1463 = .7259	995/1549 = .6423	.0836
8	1065/1463 = .7280	989/1549 = .6384	.0896
9	1067/1463 = .7293	979/1549 = .6320	.0973
10	1084/1463 = .7409	970/1549 = .6262	.1147

Results for Objective 4. After 10 splits, the MAIDDA procedure correctly predicted 241 more subjects for the Class of 1983 and 186 more subjects for the Class of 1984. This computes to a 16.71% increase in prediction for the Class of 1983 and a 12.01% increase for the Class of 1984. The shrinkage after 10 splits was found to be .1147. The optimum cross-validated prediction model occurred after the fourth split where MAIDDA demonstrated a 14.53% improvement over discriminant analysis. The shrinkage after four splits was only .0574. Thus, the unique contribution of MAID applied to the two given samples can be described as a .1453

increase in cross-validated hit rate after four splits with a very acceptable shrinkage of only .0574.

Interpretations

This section is divided into interpretations pretaining to model development and those concerning predicting attrition at USAFA. First, from a model development point of view, MAIDDA, as discussed in Objective 4 results, provides a substantial increase in prediction over classical two group discriminant analysis when applied to the two given samples. These results indicate that there exists a considerable amount of predictor variable interaction which was not accounted for in the ordinary discriminant model. As a result, it appears that a researcher can enhance prediction through the use of MAIDDA when; (a) sample data sets are large with numerous predictors; (b) predictor variable interactions exist; (c) a theoretical basis for interaction terms is not available and only haphazard guessing is used to develop interaction variables. Obviously, more tests concerning the use of MAIDDA are required to ascertain its full potential.

As far as predicting attrition at USASFA, with the set of given predictor variables, ordinary discriminant analysis produced a cross-validated hit rate of .5061 which is well below the base rate of .6475 for the Class of 1984. At this point, a researcher would be better off to predict success

for all subjects, which would be correct 64.75% of the time, even though all dropouts would be misclassified. The use of MAIDDA would improve upon the base rate by providing a hit rate of .6514 after four splits. Although this does not seem like a large improvement over the base rate, it is important to add that the misclassifications for MAIDDA will be spread over both categories whereas the base rate corresponds to misclassifying all of one category. See Tables 11 and 12 for a visual representation of this argument.

Table 11

Prediction Results for the Base Rate i.e. Predicting Success for all Subjects

Class of 1984

	ACTUAL		
	Dropouts	Persisters	TOTAL
PREDICTED			
Dropouts	0	0	0
Persisters	546	1003	1549
TOTAL	546	1003	1549
Percent correctly classified	0.0%	100.0%	64.75%

In addition, MAIDDA confirms that there are indeed differences between certain subgroups of subjects. For example, 80% or more of those candidates who have military parents will persist, whereas less than 60% of those without

military parents persist. Subjects without military parents and with an academic composite score (ACACMP) greater than

Table 12

Cross-validated Prediction Results for MAIDDA after Four Splits i.e. Subgroups 3, 6, 7, 8, and 9

Class of 1984

	ACTUAL		
	Dropouts	Persisters	TOTAL
Dropouts	197	191	388
PREDICTED			
Persisters	349	812	1161
TOTAL	546	1003	1549
Percent correctly classified	36.08%	80.96%	65.14%

3101 have a 6.9 probability of persisting, whereas those with a ACACMP less than 3101 have only a .51 probability of persisting. Of those mentioned above with an ACACMP $\geq$ 3101, if they are female their probability of persistence is .58; if they are male, it is .71. For those with an ACACMP $<$ 3101, a further breakdown into 2853 $\leq$ ACACMP $<$ 3101 and ACACMP $<$ 2853 produced .56 and .45 probabilities of persistence. It is the definite differences in persistence of these subgroups which create variable interactions and allow for the success of MAIDDA.

In order to demonstrate the use of MAIDDA in an institutional setting the following hypothetical scenerio is presented. If MAIDDA had been available to develop the model on the Class of 1983 and this model could have been used in the admissions phase for the Class of 1984, the following would have occurred: (a) 388 dropouts would have been predicted, 191 erroneously; (b) these 388 subjects would have been replaced by another 388 qualified candidates who are predicted to persist with 69.94% accuracy, thus 271 of the new group of 388 would persist; (c) of the new entering Class of 1984, 1083 would persist and 466 would dropout producing an attrition rate of .3008 versus the actual attrition rate of .3525. The result would be a 5.17% decrease in attrition.

CHAPTER V

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Summary

This study has attempted to make a significant contribution to the advancement of persister/dropout models and the field of statistics in the following areas:

1. The literature review has demonstrated that for large samples, multiple regression or two group discriminant analysis provides the simplest and most accurate predictions of dichotomus variables whose proportions normally range from .20 to .80 and in some cases .10 to .90. In addition, multiple regression and two group discriminant analysis were shown to be related models which can be designed to produce identical results. Since most statistical software packages utilize two group discriminant analysis for subject classification, it appears that this is the best choice of methodologies for samples such as those described above.

2. A serious limitation of most prediction models is the problem of incorporating interaction terms for numerous predictor variables. A new procedure, MAIDDA, developed in this study, combines two group discriminant analysis and a modified automatic interaction detector technique which can

easily account for interaction terms. This procedure was able to produce a 14.53% prediction improvement over classical two group discriminant analysis on the two data samples. The estimated shrinkage value of .0574 is extremely reasonable and demonstrates the generalization of this model. These results indicate that the unique contribution of MAID is substantial in the case of the two given samples described earlier.

3. In light of the success of MAIDDA as discussed above, it is evident that the modified AID procedure, MAID, is not only easily applied through the use of SAS, but it is also successful in accounting for variable interactions. The advantages of MAID over AID are: the convenience of using the validity coefficient instead of between sum of squares to determine the variable on which to split; the additional reduction in computation time and increase in predictability by leaving continuous variables as they are and computing a discriminant type cutoff point for the splitting process; and most importantly, the ability for any researcher to be able to apply automatic interaction detector techniques through a systematic fashion using available statistical software such as SAS without the lengthy and involved AID-IV algorithm.

4. The failure of previous prediction studies to actually classify subjects and report a hit rate is disturbing. This study has demonstrated the usefulness of the hit rate and has pointed out the dependency of χ^2 , and R^2

values on sample size. The results shown in Table 13 further support the use of the hit rate. The confusing part of χ^2 and ϕ is found after the second split, where their values are about the same for the unsplit data; however, the hit rate is improved by 170/1463 = .1162. In addition, RSQ is continually increasing despite the drop in hit rate after the second split. Furthermore, the increase in RSQ from the second to the third split is .014 with a corresponding hit rate increase of .0048, whereas, the RSQ and hit rate changes from the third to the fourth split are .005 and .0082. This indicates a lack of any consistent relationship between RSQ and hit rate. Since subject prediction is the objective of the model, it appears that the hit rate is the only reasonable measure of efficiency. In order for future models to be more accurately compared, it is hoped that researchers will conduct the actual subject classifications and report a hit rate instead of relying on other measures of model efficiency.

Table 13

χ^2, ϕ, RSQ and Hit Rate Comparisons for the First Four Splits on the Class of 1983 Sample

Split #	χ^2	ϕ	RSQ	Hit Rate
0	144.617	.314	---	843/1463 = .5762
1	159.646	.330	.091	1022/1463 = .6986
2	143.970	.315	.120	1013/1463 = .6924
3	156.437	.327	.134	1021/1463 = .6979
4	181.570	.352	.139	1033/1463 = .7088

5. The usefulness of the hit rate is demonstrated in the Chapter IV interpretations where the prediction process is hypothetically applied in an institutional setting. The fact that attrition could be reduced by 5.17% even without the SVIB variables that Dempsey and Fast (1976) found significant leads this researcher to believe that MAIDDA has demonstrated its potential to aid in the reduction of attrition. Furthermore, MAIDDA, unlike most other prediction methods, provides a systematic splitting of the original data into subgroups which are mutually exclusive and distinctly different with regard to predicting the dependent variable. This ability to subgroup the data can also be very useful for interpreting interaction effects in an explanatory model.

Conclusions

The new procedure, MAIDDA, developed in this study has demonstrated superiority over classical two group discriminant analysis when conducted on the two given samples. It has also been shown to be easily accomplished with the use of available statistical software. MAIDDA's 14.53% improvement in predicting persisters and dropouts for the USAFA Class of 1984 corresponds to a potential 5.17% reduction in attrition. This improvement can quite possibly be improved with the addition of a previously successful set of predictors, the SVIB items.

In conclusion, MAIDDA has a promising potential in predicting dichotomus dependent variables from large data sets with numerous independent variables. The specific use of MAIDDA in predicting persisters and dropouts has demonstrated an advancement in the research of attrition and the study of variable interaction.

Recommendations

The new prediction procedure, MAIDDA, has demonstrated impressive results when tested on the two samples utilized in this study. Obviously, the next step in continued research in this area is further testing of MAIDDA on other large samples with 10 or more predictors. Examining subgroup predictions more closely could reveal cases where discriminant analysis is not an improvement over predicting all successes or all failures for subjects in that subgroup. This variant to the MAIDDA procedure could provide enhanced hit rates and decreased shrinkage values. Further investigation into the splitting process should also be studied. For example, it is possible that for two competing subgroups j and k, if $TSS_j > TSS_k$ but $(r^2_{yX_{sj}} * TSS_j) < (r^2_{yX_{sk}} * TSS_k)$, then it might be more productive to split on subgroup k instead of j. In addition to the model development recommendations described above, it is clear that further study at USAFA is in order with the inclusion of the SVIB items as predictor variables. Once the Academy has developed a more sound data

base, further study in this area could possibly make a significant impact on admissions policies and, in turn, the attrition rate.

APPENDIX A

MAIDDA PREDICTION RESULTS
FOR SUBGROUPS 2 - 21

Table 14

Prediction Results for Subgroup 2

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	256	151	407
PREDICTED			
Persisters	276	523	799
TOTAL	532	674	1206
Percent Correctly Classified	48.12%	77.60%	64.60%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	194	191	385
PREDICTED			
Persisters	283	539	822
TOTAL	477	730	1207
Percent Correctly Classified	40.67%	73.84%	60.73%

Table 15

Prediction Results for Subgroup 3

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	4	3	7
PREDICTED			
Persisters	11	239	250
TOTAL	15	242	257
Percent Correctly Classified	26.67%	98.76%	94.55%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	4	14	18
PREDICTED			
Persisters	65	259	324
TOTAL	69	273	342
Percent Correctly Classified	5.80%	94.87%	76.90%

Table 16

Prediction Results for Subgroup 4

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	218	123	341
PREDICTED			
Persisters	95	149	244
TOTAL	313	272	585
Percent Correctly Classified	69.65%	54.78%	62.74%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	187	161	348
PREDICTED			
Persisters	101	145	246
TOTAL	288	306	594
Percent Correctly Classified	64.93%	47.39%	55.89%

Table 17

Prediction Results for Subgroup 5

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	30	29	59
PREDICTED			
Persisters	189	373	562
TOTAL	219	402	621
Percent Correctly Classified	13.70%	92.79%	64.90%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	26	42	68
PREDICTED			
Persisters	163	382	545
TOTAL	189	424	613
Percent Correctly Classified	13.76%	90.09%	66.56%

Table 18

Prediction Results for Subgroup 6

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	156	74	230
PREDICTED			
Persisters	19	28	47
TOTAL	175	102	277
Percent Correctly Classified	89.14%	27.45%	66.43%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	117	78	195
PREDICTED			
Persisters	16	33	49
TOTAL	133	111	244
Percent Correctly Classified	87.97%	29.73%	61.48%

Table 19

Prediction Results for Subgroup 7

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	64	43	107
PREDICTED			
Persisters	74	127	201
TOTAL	138	170	308
Percent Correctly Classified	46.38%	74.71%	62.01%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	55	69	124
PREDICTED			
Persisters	100	126	226
TOTAL	155	195	350
Percent Correctly Classified	35.48%	64.62%	51.71%

Table 20

Prediction Results for Subgroup 8

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	12	9	21
PREDICTED			
Persisters	169	353	522
TOTAL	181	362	543
Percent Correctly Classified	6.63%	97.51%	67.22%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	6	10	16
PREDICTED			
Persisters	146	362	508
TOTAL	152	372	524
Percent Correctly Classified	3.95%	97.31%	70.23%

Table 21

Prediction Results for Subgroup 9

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	27	13	40
PREDICTED			
Persisters	11	27	38
TOTAL	38	40	78
Percent Correctly Classified	71.05%	67.50%	69.23%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	15	20	35
PREDICTED			
Persisters	22	32	54
TOTAL	37	52	89
Percent Correctly Classified	40.54%	61.54%	52.81%

Table 22

Prediction Results for Subgroup 10

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	2	3	5
PREDICTED			
Persisters	100	221	321
TOTAL	102	224	326
Percent Correctly Classified	1.96%	98.66%	68.40%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	3	4	7
PREDICTED			
Persisters	86	207	293
TOTAL	89	211	300
Percent Correctly Classified	3.37%	98.10%	70.00%

Table 23

Prediction Results for Subgroup 11

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
PREDICTED Dropouts	22	14	36
PREDICTED Persisters	57	124	181
TOTAL	79	138	217
Percent Correctly Classified	27.85%	89.86%	67.28%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
PREDICTED Dropouts	11	27	38
PREDICTED Persisters	52	134	186
TOTAL	63	161	224
Percent Correctly Classified	17.49%	83.23%	64.73%

Table 24

Prediction Results for Subgroup 12

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	34	16	50
PREDICTED			
Persisters	35	83	118
TOTAL	69	99	168
Percent Correctly Classified	49.28%	83.84%	69.64%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	44	31	75
PREDICTED			
Persisters	54	78	132
TOTAL	98	109	207
Percent Correctly Classified	44.90%	71.56%	58.94%

Table 25

Prediction Results for Subgroup 13

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	41	24	65
PREDICTED			
Persisters	28	47	75
TOTAL	69	71	140
Percent Correctly Classified	59.42%	66.20%	62.86%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	19	41	60
PREDICTED			
Persisters	38	45	83
TOTAL	57	86	143
Percent Correctly Classified	33.33%	52.33%	44.76%

AD-A151 877 PREDICTING COLLEGE DROPOUTS BY COMBINING AUTOMATIC INTERACTION DETECTOR A.. (U) AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH S R SCHMIDT DEC 84 2/2

UNCLASSIFIED AFIT/CI/NR-85-25D F/G 12/1 NL

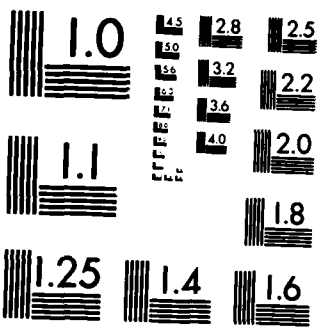


Table 26

Prediction Results for Subgroup 14

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
PREDICTED Dropouts	6	6	12
PREDICTED Persisters	76	157	233
TOTAL	82	163	245
Percent Correctly Classified	7.32%	96.32%	66.53%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
PREDICTED Dropouts	4	11	15
PREDICTED Persisters	60	141	201
TOTAL	64	152	216
Percent Correctly Classified	6.25%	92.76%	67.13%

Table 27

Prediction Results for Subgroup 15

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	10	4	14
PREDICTED			
Persisters	10	57	67
TOTAL	20	61	81
Percent Correctly Classified	50.00%	93.44%	82.72%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	5	5	10
PREDICTED			
Persisters	20	54	74
TOTAL	25	59	84
Percent Correctly Classified	20.00%	91.53%	70.24%

Table 28

Prediction Results for Subgroup 16

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	50	11	61
PREDICTED			
Persisters	0	6	6
TOTAL	50	17	67
Percent Correctly Classified	100.00%	35.29%	83.58%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	28	18	46
PREDICTED			
Persisters	5	3	8
TOTAL	33	21	54
Percent Correctly Classified	84.85%	14.24%	57.41%

Table 29

Prediction Results for Subgroup 17

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	106	60	166
PREDICTED			
Persisters	19	25	44
TOTAL	125	85	210
Percent Correctly Classified	84.80%	29.41%	62.38%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	78	55	133
PREDICTED			
Persisters	22	35	57
TOTAL	100	90	190
Percent Correctly Classified	78.00%	38.84%	59.47%

Table 30

Prediction Results for Subgroup 18

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	10	9	19
PREDICTED			
Persisters	46	88	134
TOTAL	56	97	153
Percent Correctly Classified	17.86%	90.72%	64.05%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	7	18	25
PREDICTED			
Persisters	28	81	109
TOTAL	35	99	134
Percent Correctly Classified	20.00%	81.82%	65.67%

Table 31

Prediction Results for Subgroup 19

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	4	3	7
PREDICTED			
Persisters	22	63	85
TOTAL	26	66	92
Percent Correctly Classified	15.38%	95.45%	72.83%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	2	8	10
PREDICTED			
Persisters	27	45	72
TOTAL	29	53	82
Percent Correctly Classified	6.90%	84.91%	57.32%

Table 32

Prediction Results for Subgroup 20

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	55	22	77
PREDICTED			
Persisters	13	20	33
TOTAL	68	42	110
Percent Correctly Classified	80.88%	47.62%	68.18%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	41	27	68
PREDICTED			
Persisters	8	17	25
TOTAL	49	44	93
Percent Correctly Classified	83.67%	38.64%	62.37%

Table 33

Prediction Results for Subgroup 21

Class of 1983

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	47	17	64
PREDICTED			
Persisters	10	26	36
TOTAL	57	43	100
Percent Correctly Classified	82.46%	60.47%	73.00%

Class of 1984

	ACTUAL Dropouts	Persisters	TOTAL
Dropouts	26	26	52
PREDICTED			
Persisters	25	20	45
TOTAL	51	46	97
Percent Correctly Classified	50.98%	43.48%	47.42%

BIBLIOGRAPHY

Allen, M. J. and Yen, W. M. (1979) Introduction to Measurement Theory, Brooks/Cole Publishing Company, Monterey, CA.

Anastasi, Schneiders & Meade, (1960) The Validation of a Biographical Inventory as a Predictor of College Success, College Entrance Board, (p. 1). New York.

Astin, A. W. (1972) "College Dropouts: A National Profile." (ACE Research Reports, Vol. 7, No. 1). American Council on Education.

Astin, A. W. (1973) "Student Persistence: Some Stay, Some Don't - Why?." College and University, (Vol. 48, pp. 298-306).

Astin, A. W. (1965) "Who Goes Where To College?" Science Research Association, Chicago, IL.

Astin, A. W. (1971) Predicting Academic Performance in College, The Free Press, New York.

Bean, A. G., & Covert, R. W. (1973) "Prediction of College Persistance, Withdrawal, and Academic Dismissal; A Discriminant Analysis," Educational and Psychological Measurement, (Vol 33). (pp. 407-411).

Berdie, R. F., (1962) Who Goes to College, University of Minnesota Press, Minneapolis.

Birnbaum, A. and Maxwell, A. E. (1960) "Classification based on Baye's Formula". Applied Statistics, (Vol. 9). (pp. 152-169).

Blanchi, J. R. (1980) "The Prediction of Voluntary Withdrawals from College: An Unsolved Problem" Journal of Experimental Education (Vol. 49). (pp. 29-33).

Bonner, L. W. (1956) "Factors Associated with the Academic Achievement of Freshman Students at a Southern Agricultural College." Unpublished dissertation, Penn State University. (pp. 91-92).

Boyce, R. W. & Paxson, R. C. (1965) "The Predictive Validity of Eleven Test at One State College", Educational and Psychological Measurement XXV. (pp. 1143-1147).

Boyer, R. A. (1956) "A Study of the Academic Success of Undergraduate Students Identified by Aptitude Profile: Indiana University, 1956", Unpublished dissertation, Indiana University, (pp. 99-130).

Chahbazi, P. (1957 October) "Analysis of Cornell Orientation Inventory Items on Study Habits and Their Relative Value in Prediction of College Achievement", Journal of Educational Research. (Vol. 51). (p. 119).

Comptroller General's Report the the Congress, (1976 March) "Student Attrition at the Five Federal Service Academies" Departments of Defense, Commerce, and Transportation.

Creager, J. A. & Miller, R. A. (1961) "Summary of Regression Analysis in the Prediction of Leadership Criteria , Air Force Academy Classes of 1961 through 1963" ASD-TN-61-41, Personnel Laboratory, Lackland AFB, San Antonio, TX.

Dempsey, J. R. & Fast, J. C. (1976) "Predicting Attrition: An Emphirical Study at the United States Air Force Academy", Defense Documentation Center, Cameron Station, Alexandria, VA.

DiVesta, F. J., Woodruff, A. D., & Hertel, J. P. (1949) "Motivation as a Predictor of College Success", Journal of Education and Phychological Measurement. (Vol. 9). (p. 340).

Endler, N. S. & Steinberg, D., (1963 April) "Prediction of Academic Achievement at the University Level", The Personnel and Guidance Journal. (pp. 693-699).

Fisher, R. A. (1958) Statistical Methods for Research Workers (13th Ed) New York. Hafner.

Fishman, J. A. & Pasanella, A. K. (1960) "College Admission-Selection Studies",Review of Educational Research. (Vol. 30). (pp. 298-310).

Fricke, B. G., (1956) "Prediction, Selection, Mortality and Quality Control" College and University XXXII. (pp. 34-52).

Gallant, T. F. (1965) "Academic Achievement of College Freshman and its Relationship to Selected Aspects of the Students Background", unpublished doctoral dissertation, Western Reserve University.

Garrett, H. F. (1949) "A Review and Interpretation of Investigations of Factors Related to Scholastic Success in Colleges of Arts and Sciences and Teachers Colleges", Journal of Experimental Education, XVIII (pp. 91-138).

Goodman, L. A. (1976) "The Relationship Between Modified and Usual Multiple-Regression Approaches to the Analysis of Dichotomous Variables", Sociology Methodology, (pp. 83-110).

Hays, W. L. (1963) Statistics, New York, Holt, Rhinehart and Winston.

Jensen, H. (1983) Office of Institution Research, USAF Academy, CO. Personal Interview.

Jernigan, E. E. (1969) "A Statistical Analysis of Three Groups of Students to Determine if Selected Scores are Related to Certain Measures of Success for each of the Four Years of Education at the Air Force Academy", Dissertation University of Denver.

Karathanos, D. (1975) An Adaption of the Automatic Interaction Detector - Version 4 (AID 4) Procedure, PhD Dissertation, University of Northern Colorado, College of Education, Greeley, CO.

Knoke, D. (1975 May) "A Comparison of Log-linear and Regression Models for Systems of Dichotomous Variables", Sociological Methods & Research. (Vol. 3, pp. 416-434).

Lavin, D. E. (1967) The Prediction of Academic Performance by John Wiley & Sons, New York.

Marascuilo, L. A. & Levin, J. R. (1983) Multivariate Statistics in the Social Sciences: A Researchers Guide, Brooks/Cole Publishing Company, Monterey, CA.

Michael, W. B. and Jones, R. A. (1962) "High School Records and College Board Scores as Predictors of Success in a Liberal Arts Program During Freshman Year of College", Educational and Psychological Measurement, XXII. (pp. 399-400).

Michael, W. B. & Perry, N. C. (1956) "The Comparability of the Simple Discriminant Function and Multiple Regression Techniques." Journal of Experimental Education, (Vol. 24. pp. 297-301).

Miller, R. E. (1966) "Predicting First Year Achievement of Air Force Academy Cadets, Class of 1967", PRL-TR-66-18, AD-660-121. (Lackland AFB, Texas: Personnel Research Laboratory, Aerospace Medical Division).

Miller, R. E. (1968) "Predicting First Year Achievement of Air Force Academy Cadets, Class of 1968", AFHRL-TR-68-103, (Lackland AFB, Texas: Personnel Research Laboratory, Air Force Systems Command).

Myers, R. C., & Schultz, D. G., (1950) "Predicting Academic Achievement with a New Attitude Interest Questionaire - I", Journal of Educational and Psychological Measurement, (Vol. 10, pp. 654-655).

Nerlove, M. & Press, S. J. (1973) "Universate and Multivariate Log-linear and Logistic Models", Rand Corporation Santa Monica, CA.

Panos, R. J. & Astin, A. W. (1968) "Attrition Among College Students", American Educational Research Journal, (Vol. 5).

Pantages, T. J. & Creddon, C. F. (1978) "Studies of College Attrition 1950-1975", Review of Educational Research, (Vol. 48, pp. 49-101).

Pascarella, E. T. & Chapmen, D. W. (1983) "Validation of a Theoretical Model of College Withdrawal: Interaction Effects in a Multi-Institutional Sample." Research in Higher Education, (Vol. 19, No. 1).

Pascarella, E. T. & Chapman, D. W. (1983) "A Multi-institutional Path Analytic Validation of Tinto's Model of College Withdrawal", American Educational Research Journal, (Vol. 20, No. 1, pp. 87-102).

Pascarella, E. T., Duby, P. B., Miller, V. A. & Rasher, S.P. (1981) "PreEnrollment Variables and Academic Performance as Predictors of Freshman Year Persistence, Early Withdrawal, and Stopout Behavior in an Urban, Non-Residential University", Research in Higher Education, (Vol. 15, No. 4).

Pedhazer, E. J. (1982) Multiple Regression in Behavioral Research, Holt, Rinehart and Winston.

Rose, H. A. & Elton, C. F. (1966) "Another Look at the College Dropout", Journal of Counseling Psychology, (Vol. 13, pp. 242-245).

Rugg, E. A. (1983) (1983) "Design and analysis Considerations for Longitudinal Retention and Attrition Studies", College and University. (Vol. 58, No. 2, pp. 119-134).

SAS User's Guide: Statistics, 1982 Edition, SAS Institute, Cary, NC.

Smith, A. T (1982) "Cross Validation and Discriminative Analysis Techniques in a College Student Attrition Application", College Student Journal, (Vol. 16, p. 34).

Sonquist, J. A., Baker, E. L., & Morgan, J. N. (1971) Searching for Structure (ALIAS - AID-III) Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, MI.

Summerskill, J. (1962) "Dropouts from College", In N. Stanford (Ed.) The American College, New York. John Wiley.

Tatsuoka, M. M. (1971) Multivariate Analysis: Techniques for Educational and Psychological Research, John Wiley and Sons, New York.

Twining, C. W. (1957) "The Relationship of Extra-curricular Activities to School Marks", School Activities, (Vol. 28, pp. 181-184).

United States Air Force Academy Catalog 1981-82.

Wasserman, M. (1979) "An Evaluation of a Compensatory Introductory Sociology Section", Journal of Experimental Education. (Vol. 47, pp. 162-171).

Webb, S. C. & McCall, J. N. (1963) "Predictors of Freshman Grades in a Southern University", Education and Psychological Measurement , XIII. (pp. 660-663).

VITA

NAME: Stephen R. Schmidt

BIRTH: 30 March 1948
Fredericksburg, Texas 78624

FAMILY: Judy B. (Hopmann) Schmidt, wife
Jeffrey Garrett Schmidt, son

EDUCATION: B.S. in Mathematics, 1970
United States Air Force Academy

M.S. in Operations Research, 1980
University of Texas at Austin

EXPERIENCE: Instructor of Mathematics 1980-1982
United States Air Force Academy

END

FILMED

4-85

DTIC